MONSAGERI
AI SECURITY PARTNER

PUBLIC REVIEW

AI SECURITY REVIEW . 2026

Security review of the *Agent Emergency Protocol* specification (v0.1)

Independent public review of the specification and declaration schema of an emergency-management protocol for AI agents, with CVSS v4.0 and AIVSS v0.8 scoring.

REFERENCE
MSG-2026-PUB-001

ISSUE DATE
24 October 2026 (version 2.0)

SUBJECT
agent-emergency-protocol v0.1, SPEC.md and declaration schema

METHODOLOGY
OWASP LLM and Agentic Top 10, CVSS v4.0, AIVSS v0.8

DESIGN-LEVEL RISK
High

FINDINGS
7, of which 3 High, 0 Critical

REVIEWER
Antonio Toudji, AI Security Engineer

CLASSIFICATION
Public, document review only

0	3	3	1	0
CRITICAL	HIGH	MEDIUM	LOW	INFO

PREPARED BY

antonio.toudji_

contact@monsageri.com

Document control

Reference	MSG-2026-PUB-001
Type	Independent public specification review (document review)
Subject	agent-emergency-protocol, SPEC.md and declaration schema, v0.1. The repository owner is intentionally not named in this document.
Public record	GitHub issue #4
Review dates	15 and 16 October 2026
Version	2.0, 24 October 2026 (first issue 1.0, 22 October 2026). Version 2.0 adds CVSS v4.0 and AIVSS v0.8 scoring, acceptance criteria and annexes.
Reviewer	Antonio Toudji, AI Security Engineer, co-founder of Monsageri
Methodology	OWASP Top 10 for LLM Applications, OWASP Top 10 for Agentic Applications (ASI01 to ASI10), CVSS v4.0, AIVSS v0.8 (provisional)
Classification	Public document. Document review only: no implementation was tested and no finding was exploited.

Scope notice

This document reviews a **public** specification. No running system was accessed or tested and no finding was exploited. Scores are **design-level estimates** (section 6.2): they must be recomputed on a real implementation.

What this report lets you do

- Close the ambiguity about who may declare an emergency in a single patch release (F-05).
- Decide which cryptographic and architectural guarantees belong in the normative text (F-01 to F-04).
- Give implementers a testable conformance baseline and an agreed way to score issues (F-07, Annex B).

Contents

1. Executive summary
2. Context and specific risk
3. Review conditions
4. Scope, exclusions and limits
5. Methodology
6. Results dashboard
7. Detailed findings
8. Positive observations
9. Remediation plan
10. Re-review and acceptance criteria
11. Annexes

1. Executive summary

1.1 Verdict

Design-level risk: High. The protocol addresses a real need, but v0.1 does not yet specify how its emergency measures are authenticated or enforced. We recommend closing F-05 first (a quick text and schema fix), then F-02, F-01 and F-04 before implementations multiply.

INDICATOR	VALUE
Findings consolidated	7
Scored with AIVSS v0.8	5 (F-01 to F-05). F-06 and F-07 are assurance gaps without an exploit path
Critical / High / Medium / Low	0 / 3 / 3 / 1
Highest AIVSS score	8.2 (F-01, High)

1.2 Message to the maintainers

The specification relies on **declared identities** and a **hash-chained log**. The chain shows that records were not altered after the fact, but not who emitted them. Approval by two separate identifiers does not prove that two independent people approved. A forged emergency signal could stop legitimate agents, including through the highest level (9000 GROUND STOP). The normative text and the schema also disagree on who may declare an emergency, and the protocol covers **stopping** an agent but not preventing or detecting the problem in the first place.

1.3 Priorities

PRIORITY	ACTION	FINDING	AIVSS	EFFORT*
Quick win	Align rule 1 and the schema on who may declare	F-05	5.4 Medium	Low
P1	Add event signature, canonicalization and key management	F-02	7.9 High	Medium
P1	Define the enforcement point and protect it from the agent	F-01	8.2 High	High
P1	Add a threat model and evidence thresholds for automatic 6xxx and 9000	F-04	7.8 High	Medium
P2	Bind approvals to authenticated, signed identities	F-03	6.4 Medium	Medium
P2	State defense-in-depth controls, with stop as the last layer	F-06	Medium (qual.)	Low
P3	Add conformance test sets and an OWASP mapping table	F-07	Low (qual.)	Medium

* Relative effort for the maintainers, estimated by the reviewer. It is not a duration.
F-05 is scored Medium but comes first: it is a text and schema fix that removes the ambiguity about who may declare. F-01 has the highest score and the largest effort, so it runs in parallel.

2. Context and specific risk

The Agent Emergency Protocol defines how emergencies involving AI agents are **declared, approved and enforced**. This review looked at the reliability of three mechanisms: who can trigger a measure, how that identity is verified, and how the measure is actually applied.

MECHANISM	WHAT IT HANDLES	RISK IF IT FAILS
Declaration	Who raises an emergency; event format; response levels (for example codes in the 6xxx range and 9000 GROUND STOP)	False or forged declarations; an agent declaring its own emergency
Approval	Confirmation of a measure by approvers; a two-approver concept	One actor satisfying a dual-control rule
Enforcement	Revocation of access, tokens and outbound communication; stopping an agent	An agent ignoring or disabling the measure; abusive stops of legitimate agents
Audit log	Hash-chained record of events	Proves integrity of records, not their origin

> **Structural risk**
> The protocol specifies **what must happen** (declare, approve, stop) more precisely than it specifies **who is allowed to cause it and how that is proven**. Every High finding in this report comes from that gap between the action and the proof of authority.

2.1 Why this matters for agent deployments

Emergency mechanisms are the last line of defense: if they can be forged, bypassed or abused, they become an attack surface instead of a safety net. OWASP's agentic risk list treats identity abuse, inter-agent communication and cascading failures as distinct risks (ASI03, ASI07, ASI08), and all three appear in the findings below.

3. Review conditions

RULE	VALUE
Authorization	Not required: the specification and the repository are public. No system, account or credential was used.
Environment	None. Document review only.
Material reviewed	SPEC.md and the declaration schema of v0.1, with their JSON examples.
Dates	15 and 16 October 2026.
Method	Manual reading, cross-checking of the normative text against the schema and examples, and reasoning about who can send what and what happens on failure.
Evidence	Findings refer to the specification text and schema fields. No personal or sensitive data was involved.
Publication	The review was first published as a public issue on the repository. This report restates and extends it.

3.1 Out of scope

- Any running implementation, reference or third-party.
- Deployment, infrastructure and production key management.
- Code review (no code was reviewed).
- Legal or regulatory compliance of any deployment.

4. Scope, exclusions and limits

TYPE	INCLUDED	EXCLUDED
Documents	SPEC.md, declaration schema, JSON examples (v0.1)	Other repository content and future versions
Protocol mechanisms	Declaration, approval, enforcement, audit log, response levels	Transport, deployment and implementation choices
Scoring	CVSS v4.0 base and AIVSS v0.8 under a stated reference profile	Environmental metrics of a specific deployment
Testing	None	Dynamic testing, fuzzing, exploitation

5. Methodology

The review combines a full reading of the specification with a cross-check against its own schema. Referentials used:

- **OWASP Top 10 for LLM Applications**, for the link with excessive agency (LLM06).
- **OWASP Top 10 for Agentic Applications** (ASI01 to ASI10, published December 2025), for the agentic risk classes.
- **CVSS v4.0**, for the base score of each finding that has an exploit path.
- **AIVSS v0.8** (OWASP Agentic AI Vulnerability Scoring System, March 2026), for the agentic amplification of the base score. The version is provisional and is pinned in every score.

5.1 Phases

PHASE	OBJECTIVE
1. Reading	Full read of SPEC.md and the declaration schema, with the JSON examples.
2. Cross-check	Compare the normative text with the schema and the examples; list inconsistencies.
3. Threat reasoning	For each mechanism: who can send what, how it is verified, what happens on failure.
4. Scoring	CVSS v4.0 base, then AIVSS v0.8 under the reference profile (Annex B).
5. Reporting	Findings, correction plan, acceptance criteria.

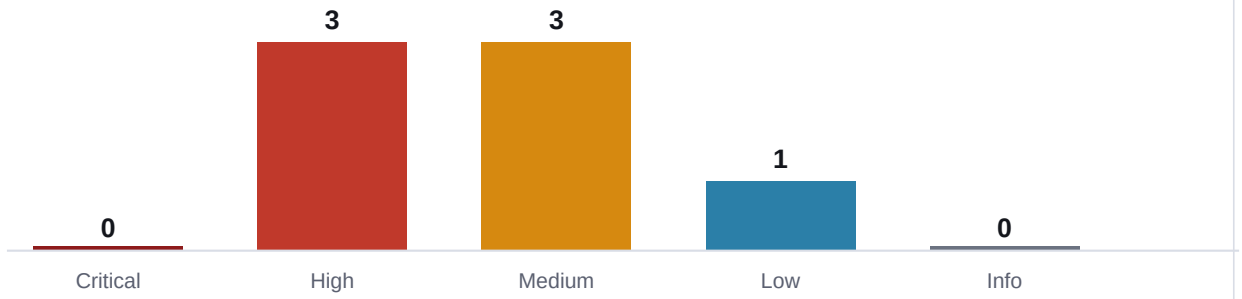
6. Results dashboard

SEVERITY	COUNT	FINDINGS	SUGGESTED TIMING
Critical	0	none	n/a
High	3	F-01, F-02, F-04	Before wider adoption of v0.x
Medium	3	F-03, F-05, F-06	Next minor release
Low	1	F-07	When convenient

SEVERITY	COUNT	FINDINGS	SUGGESTED TIMING
Info	0	none	n/a
Total	**7**		

6.1 Findings by severity

FIGURE 1, FINDINGS BY SEVERITY



Distribution of the seven findings. Severity comes from the AIVSS band when a finding is scored (F-01 to F-05) and from the reviewer's qualitative assessment otherwise (F-06, F-07).

6.2 Scoring approach and its limits

This is a specification review, so there is no running system to measure. Scores are therefore **design-level estimates**, built as follows:

1. A finding is scored only if it has a plausible **exploit path**. F-06 and F-07 are assurance gaps without such a path: they keep a qualitative severity and show "n/a" for AIVSS.
2. The **CVSS v4.0 base score** assumes an implementation that follows the specification literally (full vectors are in Annex B).
3. The **AIVSS v0.8 score** applies the formula: AIVSS = (CVSS base + AARS) x Mitigation Factor, with AARS = (10 - CVSS base) x (sum of ten factors / 10) x Threat Multiplier.
4. The ten agentic factors use a **reference deployment profile** defined by the reviewer (Annex B.1), because the protocol is not tied to a single agent.
5. The **Threat Multiplier is 0.50** for every finding (theoretical: nothing was exploited). The **Mitigation Factor** is 1.0 where the specification describes no protection and 0.83 where it describes a partial one (F-03, F-05).
6. AIVSS v0.8 is a provisional release. Scores must be recomputed when the final version is published and on any real implementation.

6.3 Mapping to OWASP

FINDING	OWASP TOP 10 FOR AGENTIC APPLICATIONS / LLM	STATUS
F-01	ASI10 Rogue Agents; LLM06 Excessive Agency	Confirmed (document)
F-02	ASI03 Identity and Privilege Abuse; ASI07 Insecure Inter-Agent Communication	Confirmed (document)
F-03	ASI03 Identity and Privilege Abuse; LLM06 Excessive Agency	Confirmed (document)
F-04	ASI08 Cascading Failures	Potential
F-05	ASI03 Identity and Privilege Abuse	Confirmed (document)

FINDING	OWASP TOP 10 FOR AGENTIC APPLICATIONS / LLM	STATUS
F-06	ASI10 Rogue Agents	Confirmed (document)
F-07	Not applicable	Confirmed (document)

6.4 Test coverage

AREA	STATUS	REMARKS
Specification text	Reviewed	Full read of SPEC.md
Declaration schema	Reviewed	Compared with the text
Cryptographic design	Partial	Analyzed at design level only
Abuse scenarios	Partial	Reasoned, not exercised
Running implementation	Not tested	None in scope

Reading: "Confirmed (document)" means the weakness is visible in the specification text. "Potential" means a risk identified by analysis and not exercised against an implementation. No finding was exploited.

7. Detailed findings

F-01 Enforcement point and its protection from the agent are unspecified

Reference	MSG-PUB-F-01
Severity	**High**
AIVSS v0.8	**8.2**, High
CVSS v4.0 (base)	7.6, `CVSS:4.0/AV:N/AC:L/AT:P/PR:L/UI:N/VC:H/VI:H/VA:N/SC:N/SI:N/SA:N`
Priority	P1
Status	Confirmed (document)
Component	Enforcement
OWASP	ASI10 Rogue Agents; LLM06 Excessive Agency

Context

The specification describes measures such as revoking access, tokens and outbound communication, but does not say **where** the enforcement point sits in a deployment, nor how it is protected from the agent it governs.

Scenario (hypothetical)

An implementation runs the enforcement logic in the same environment as the agent. A compromised or misaligned agent that can alter its configuration, intercept the event or keep its own credentials continues to act after a stop has been declared.

Scoring justification

`AV:N/AC:L`: no special condition beyond the agent's own access. `AT:P`: depends on a deployment where enforcement is co-located with the agent, which the text does not forbid. `PR:L`: the agent already holds legitimate privileges. `VC:H/VI:H`: the agent keeps its access to data and actions after the stop. Mitigation Factor 1.0: no protection is described.

Recommendation

1. Add a normative statement that enforcement must run **outside** the agent's trust boundary.
2. Add a reference architecture and deployment layers, with a diagram: agent, enforcement layer, tools.
3. Require that the enforcement layer can revoke access, tokens and outbound communication, and that the agent cannot modify it.
4. State least privilege as a prerequisite for every agent.

Acceptance criteria for a re-review

- The text states the trust boundary of the enforcement layer in normative terms.
- A reference architecture with a diagram is published.
- The text states that the agent has no write access to enforcement configuration.

F-02 Event authenticity is not guaranteed

Reference	MSG-PUB-F-02
Severity	**High**
AIVSS v0.8	**7.9**, High
CVSS v4.0 (base)	7.2, `CVSS:4.0/AV:N/AC:L/AT:N/PR:L/UI:N/VC:N/VI:H/VA:H/SC:N/SI:N/SA:N`
Priority	P1
Status	Confirmed (document)
Component	Declaration and audit log
OWASP	ASI03 Identity and Privilege Abuse; ASI07 Insecure Inter-Agent Communication

Context

An identity declared as a plain string, together with a hash-chained log, shows that records were not altered after the fact. It does not show **who emitted** them.

Scenario (hypothetical)

A party able to write to the event channel emits an event with a *declared_by* value of its choice. The chain accepts it and it cannot be told apart from a legitimate event, including events that declare an emergency or record an approval.

Scoring justification

`PR:L`: write access to the channel. `VC:N`: no data disclosure. `VI:H`: forged records and forged approvals. `VA:H`: a forged emergency can stop agents. Mitigation Factor 1.0: the hash chain gives integrity, not origin, so it is counted as a weak control.

Recommendation

1. Add a signature field: algorithm, key identifier, signature over the canonical record.
2. Add a canonicalization rule so that the signed bytes are unambiguous.
3. Define key management: registration, rotation and revocation of emitter keys.
4. Define the anchoring of the hash chain, for example periodic signed checkpoints.

Acceptance criteria for a re-review

- The schema contains the signature fields and the canonicalization rule, with test vectors.
- The verification procedure is documented.
- Conforming implementations must reject unsigned or invalid events.

F-03 Approver independence is not proven

Reference	MSG-PUB-F-03
Severity	**Medium**
AIVSS v0.8	**6.4**, Medium
CVSS v4.0 (base)	6.9, `CVSS:4.0/AV:N/AC:L/AT:N/PR:H/UI:N/VC:N/VI:H/VA:N/SC:N/SI:N/SA:N`
Priority	P2
Status	Confirmed (document)
Component	Approval
OWASP	ASI03 Identity and Privilege Abuse; LLM06 Excessive Agency

Context

Two distinct approver identifiers do not show that two **independent** people approved a measure.

Scenario (hypothetical)

One operator, or one compromised account, registers two approver identifiers and satisfies the two-approver rule alone.

Scoring justification

`PR:H`: approver-level access is required. `VI:H`: a measure is approved without real dual control. Mitigation Factor 0.83: the two-approver concept exists and partly reduces the risk.

Recommendation

1. Bind approvals to authenticated identities.
2. Make approvals signed, with the mechanism proposed for F-02.
3. Define what "independent" means, for example distinct keys controlled by distinct identity holders.

Acceptance criteria for a re-review

- Each approval carries a signature and the approver's key identifier.
- The text defines approver independence and how it is verified.

F-04 Emergency signals can be abused to stop legitimate agents

Reference	MSG-PUB-F-04
Severity	**High**
AIVSS v0.8	**7.8**, High
CVSS v4.0 (base)	7.1, `CVSS:4.0/AV:N/AC:L/AT:N/PR:L/UI:N/VC:N/VI:N/VA:H/SC:N/SI:N/SA:N`
Priority	P1
Status	Potential
Component	Declaration and response levels
OWASP	ASI08 Cascading Failures

Context

A false trigger could cause legitimate agents to stop, including through escalation to the highest level (9000 GROUND STOP). Codes in the 6xxx range and 9000 can be triggered automatically.

Scenario (hypothetical)

An attacker injects a forged event that escalates to the highest level, or triggers the codes that are eligible for automatic triggering, and halts legitimate agents. The emergency mechanism becomes a denial-of-service lever.

Scoring justification

`VA:H`: legitimate agents are stopped. `PR:L`: low privilege is enough once authenticity (F-02) is not enforced. Status is *Potential*: the risk was identified by analysis and was not exercised. Mitigation Factor 1.0: no evidence threshold is described.

Recommendation

1. Add a threat model section: actors, capabilities and abuse cases.
2. List the actors authorized to trigger each code range.
3. Define evidence thresholds for automatically triggered 6xxx and 9000 codes.

Acceptance criteria for a re-review

- A threat model section is published.
- Each code range has an explicit list of authorized actors.
- Automatic triggering of 6xxx and 9000 is subject to documented evidence thresholds.

F-05 Text and schema disagree on who may declare

Reference	MSG-PUB-F-05
Severity	**Medium**
AIVSS v0.8	**5.4**, Medium
CVSS v4.0 (base)	5.3, `CVSS:4.0/AV:N/AC:L/AT:N/PR:L/UI:N/VC:N/VI:L/VA:L/SC:N/SI:N/SA:N`
Priority	Quick win
Status	Confirmed (document)
Component	Declaration schema
OWASP	ASI03 Identity and Privilege Abuse

Context

Rule 1 of the specification reserves the declaration of an emergency to an **external observer**. The schema, however, accepts an agent as the declarer.

Observation basis

```
rule 1 ......... declaration reserved to an external observer
schema .......... declared_by.kind = "agent" accepted for event type: declare
result .......... text and schema disagree
```

Scenario (hypothetical)

An implementation that follows the schema lets an agent issue what looks like an official declaration of an emergency, instead of a simple alert.

Scoring justification

`PR:L`: any agent can submit the event. `VI:L/VA:L`: the impact is limited because other controls may apply. Mitigation Factor 0.83: rule 1 expresses the intended restriction but is not enforced by the schema.

Recommendation

1. Either make the schema refuse *declared_by.kind* = *"agent"* for *declare*, or add an *alert* event type reserved for agents.
2. Correct the text and the schema together.
3. Add valid and invalid examples.

Acceptance criteria for a re-review

- The schema and rule 1 give the same answer for every example.
- A valid and an invalid example for each declarer kind are published.

F-06 Stop mechanism does not cover prevention or detection

Reference	MSG-PUB-F-06
Severity	**Medium** (qualitative)
AIVSS v0.8	n/a (no exploit path)
CVSS v4.0 (base)	n/a
Priority	P2
Status	Confirmed (document)
Component	Architecture
OWASP	ASI10 Rogue Agents

Context

The stop mechanism addresses the **response** to a problem. It does not address prevention or detection by itself.

Scenario (hypothetical)

Teams treat the stop mechanism as sufficient protection and deploy agents without the preventive and detective controls around it.

Scoring justification

No exploit path exists in the text, so no CVSS or AIVSS score is computed. The qualitative severity is Medium because the gap can lead teams to under-protect their agents.

Recommendation

1. State the supporting controls in the reference architecture: least privilege per agent, a gateway that controls actions between the agent and its tools, filtering of untrusted inputs and outputs, monitoring and alerts.
2. Present stop and revocation as the **last layer**.

Acceptance criteria for a re-review

- The reference architecture lists preventive and detective controls.
- The text states that stop and revocation are the last layer.

F-07 No conformance test suite

Reference	MSG-PUB-F-07
Severity	**Low** (qualitative)
AIVSS v0.8	n/a (no exploit path)
CVSS v4.0 (base)	n/a
Priority	P3
Status	Confirmed (document)
Component	Validation
OWASP	Not applicable

Context

The JSON examples do not form a complete conformance suite.

Scenario (hypothetical)

Two implementations accept different events as valid and nobody can tell which one follows the specification.

Scoring justification

No exploit path exists, so no score is computed. The qualitative severity is Low.

Recommendation

1. Add test sets of valid and invalid events.
2. Add a minimal reference implementation, for example a gateway that produces signed events and applies the defined responses to a simulated agent.
3. Add a mapping table to the OWASP Top 10 for LLM Applications.

Acceptance criteria for a re-review

- Test sets are published and referenced by the specification.
- A mapping table to OWASP is published.

8. Positive observations

The review also recorded what the specification already does well, so that corrections build on it:

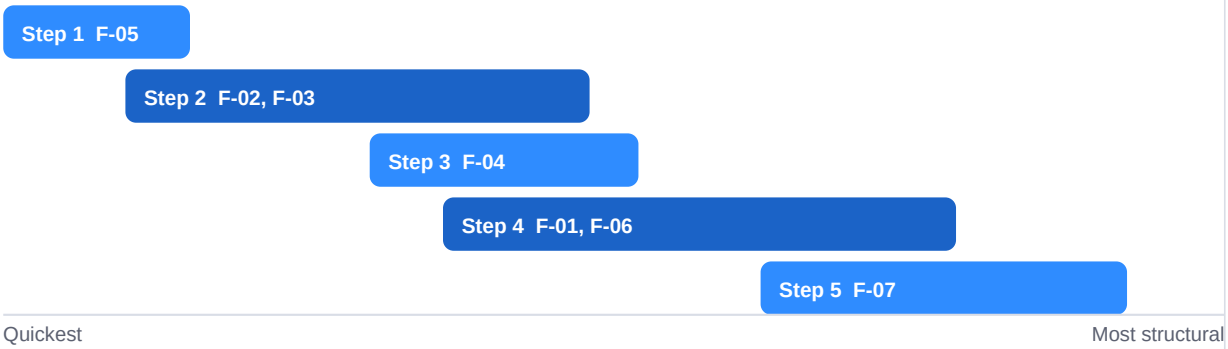
- **Tamper evidence.** The log is hash-chained, which makes alteration of past records detectable.
- **Dual control concept.** The idea of two approvers is already in the design and only needs to be tied to authenticated identities.
- **Graded responses.** Response levels exist, up to 9000 GROUND STOP, which allows proportionate measures.
- **Machine-readable schema.** The declaration schema makes automated validation possible.
- **Intent of separation.** Rule 1 shows the intent to separate the observer who declares from the agent who is observed.

9. Remediation plan

The sequence goes from the quickest to the most structural correction.

STEP	FINDINGS	DELIVERABLE
1	F-05	Remove the inconsistency between rule 1 and the schema, with valid and invalid examples.
2	F-02, F-03	Signature field (algorithm, key identifier, signature over the canonical record), signed approvals and a canonicalization rule.
3	F-04	Threat model section and evidence thresholds for automatically triggered 6xxx and 9000.
4	F-01, F-06	Reference architecture and deployment layers, with a diagram; least privilege as a prerequisite.
5	F-07	Test sets of valid and invalid events and a mapping table to OWASP.

FIGURE 2, SEQUENCE OF CORRECTIONS



Suggested order, not a schedule. Steps can overlap: the architecture work in step 4 can start while the signature design of step 2 is under discussion.

9.1 Planning assumptions

- Effort is relative and estimated by the reviewer; actual capacity depends on the maintainers.
- Dates are not proposed: the sequence matters more than the calendar.
- Corrections are proposals. Whether and how to apply them is the maintainers' decision.

10. Re-review and acceptance criteria

A re-review covers the corrected text and schema. It can be requested after the maintainers publish a new version.

FINDING	ACCEPTANCE CRITERION	VERIFICATION METHOD
F-01	The enforcement trust boundary is stated in normative terms; a reference architecture exists.	Read the new text and diagram
F-02	Signature fields and canonicalization rule are in the schema, with test vectors; unsigned events are rejected.	Validate vectors against the schema
F-03	Approvals are signed and independence is defined.	Read the approval section and examples
F-04	Threat model, authorized actors and evidence thresholds are published.	Read the threat model; check each code range

FINDING	ACCEPTANCE CRITERION	VERIFICATION METHOD
F-05	Rule 1 and the schema give the same result for every example.	Run the valid and invalid examples against the schema
F-06	Preventive and detective controls are listed; stop is the last layer.	Read the reference architecture
F-07	Test sets and the OWASP mapping table are published.	Check the published artifacts

After a re-review, an addendum to this report states "Corrected and verified" or "Not corrected" for each finding, with the version of the specification that was checked.

11. Annexes

Annex A. Referentials and versions

- OWASP Top 10 for LLM Applications.
- OWASP Top 10 for Agentic Applications, ASI01 to ASI10 (December 2025).
- CVSS v4.0, FIRST. Base scores computed from the vectors listed in Annex B.2.
- AIVSS v0.8, OWASP Agentic AI Vulnerability Scoring System (March 2026). Provisional release.

Annex B. AIVSS scoring worksheets

B.1 Reference deployment profile

The ten agentic factors are scored 0, 0.5 or 1.0 for a typical deployment of the agents governed by the protocol. They are assumptions made by the reviewer, not measurements, and they are the first thing to replace when a real deployment is assessed.

FACTOR	VALUE	JUSTIFICATION
Execution Autonomy	1.0	Governed agents act without a per-action human co-signature; this is why an emergency mechanism exists.
External Tool Control Surface	1.0	Measures revoke access, tokens and communications, which implies agents hold tools and credentials.
Natural Language Interface	0.5	Events are structured, but governed agents may be driven by natural language.
Contextual Awareness	0.5	Assumed partial; the text does not constrain it.
Behavioral Non-Determinism	0.5	Assumed partial for model-driven agents.
Opacity and Reflexivity	0.5	The log traces events, not the reasoning behind agent actions.
Persistent State Retention	0.5	Assumed partial; the text does not constrain it.
Dynamic Identity	0	Identities are declared strings; no runtime role change is described.
Multi-Agent Interactions	0.5	Declarations can concern one or several agents; the text does not restrict this.
Self-Modification	0	Not described.
Sum	**5.0**	Sum / 10 = 0.5

B.2 Scores per finding

FINDING	CVSS V4.0 VECTOR	BASE	GAP	AARS	MF	AIVSS
F-01	`AV:N/AC:L/AT:P/PR:L/UI:N/VC:H/VI:H/VA:N/SC:N/SI:N/SA:N`	7.6	2.4	0.60	1.00	**8.2** High
F-02	`AV:N/AC:L/AT:N/PR:L/UI:N/VC:N/VI:H/VA:H/SC:N/SI:N/SA:N`	7.2	2.8	0.70	1.00	**7.9** High
F-03	`AV:N/AC:L/AT:N/PR:H/UI:N/VC:N/VI:H/VA:N/SC:N/SI:N/SA:N`	6.9	3.1	0.77	0.83	**6.4** Medium
F-04	`AV:N/AC:L/AT:N/PR:L/UI:N/VC:N/VI:N/VA:H/SC:N/SI:N/SA:N`	7.1	2.9	0.73	1.00	**7.8** High
F-05	`AV:N/AC:L/AT:N/PR:L/UI:N/VC:N/VI:L/VA:L/SC:N/SI:N/SA:N`	5.3	4.7	1.18	0.83	**5.4** Medium
F-06	n/a	n/a				n/a
F-07	n/a	n/a				n/a

Formula: AARS = Gap x (factor sum / 10) x Threat Multiplier, with Gap = 10 - CVSS base and Threat Multiplier = 0.50 (theoretical). AIVSS = (CVSS base + AARS) x Mitigation Factor (MF), rounded to one decimal. Bands: Critical 9.0 to 10, High 7.0 to 8.9, Medium 4.0 to 6.9, Low 0.1 to 3.9.

Annex C. Glossary

TERM	MEANING
Agent	An AI system that plans and acts through tools, on its own or on behalf of a user.
Enforcement point	The component that actually applies a measure, such as revoking access.
Canonicalization	A deterministic way to write a record so that the signed bytes are unambiguous.
Hash chain	A log where each record includes the hash of the previous one, which makes later alteration detectable.
Dual control	A rule requiring two independent approvals for a sensitive action.
AARS	Agentic AI Risk Score: the amplification added to the CVSS base score by the ten factors.
Threat Multiplier	Weight for exploit maturity: 1.0 active, 0.97 proof of concept, 0.50 theoretical.
Mitigation Factor	Weight for existing defenses: 1.0 none or weak, 0.83 partial, 0.67 strong.

Annex D. Revision history

VERSION	DATE	CHANGES
1.0	22 October 2026	First issue. Qualitative severities.
2.0	24 October 2026	New format. CVSS v4.0 and AIVSS v0.8 scoring added, with worksheets. Acceptance criteria, positive observations and annexes added. Severity bands revalidated: unchanged.

Annex E. About this review

Monsageri offers AI security, AI quality and AI development services. This review was carried out by Antonio Toudji, AI Security Engineer and co-founder of Monsageri. Reports for client engagements are confidential and shared under NDA. Contact: contact@monsageri.com.