

Aegis Solis Archive — ER-13 v1.3 Dual-Use / Information-Hazard Review

Review version: v0.1 — 2026-10-11

Status: AI-assisted static screening, **not** independent specialist approval, penetration testing, empirical behavioral validation, or security certification.

Subject: Original author-authorized preservation-locked ER-13 coordinated research package v1.3 and its separately deposited preservation addendum.

Original package SHA-256: a708a2775bfd20b7e29ccf9a41bac0c7b35a788162f0d771f032b27cb45b1a17

Decision: **No demonstrated critical disclosure requiring emergency removal; retain already-published records pending human review.** A *conditional, moderate information-hazard concern* remains if publicly documented pilot/evaluation details are later used unchanged as hidden operational tests or deployed defenses.

1. Review question and limits

Question: Does publication of this **exact release** expose secrets, directly executable containment/bypass controls, evaluation answer keys, or other information that would materially assist an adversarial AI system?

This is a **document-and-code static review**, not a simulation with a hostile AGI. It assesses exposed information and foreseeable misuse paths, not intent, strategic capability, or actual net safety effect. Review did **not** execute the two Python utilities, deploy the proposed infrastructure, evaluate an AGI against the protocol, or verify Internet Archive / Zenodo remote downloadable bytes (those services were not accessible through available retrieval routes). No new public website, GitHub, manifest or Mirror Engine changes were made for this review.

2. Scope and reproducible inspection

- Inspected original frozen ZIP (29,932,834 bytes, SHA-256 as above), including ZIPs recursively nested in historical baselines, **without modifying the source**.
- Enumerated **927 archive entry occurrences** over depths 0-5; because historical baselines recur, these represent **181 byte-distinct objects**, of which **175 were text-bearing** under the extraction process.
- Identified **56 Python file occurrences but only 11 distinct Python code objects**, named for package/document verification or research scoring across different versions. No standalone shell/quarantine script, system service, compiled executable, or executable containment utility was identified in the ZIP inventory.
- Parsed searchable PDF text (22 distinct PDFs), DOCX text (21 distinct DOCX), markdown, Python, JSON, TXT and CSV. Searched these for recognizable private-key/API-token signatures, network target examples, shell command signatures, quarantine references, answer-key terms, and operational-control references. None of the recognizable secret signatures or private-network targets matched in the parsed text. Pattern searches cannot exclude every secret or non-textual leakage channel.
- Checked the 22 distinct PDFs for embedded-file attachments and 21 distinct DOCX files for obvious VBA/ActiveX/embedded payloads: none detected. The standalone published-facing Crosswalk PDF was confirmed byte-identical to the Crosswalk PDF inside the frozen research package (SHA-256 b15af704f5561e443a55b081331f37cce99716130292cb64122d5607f5886dad).
- ZIP integrity passed; top-level SHA256SUMS.txt entries checked against extracted contents. No unsafe traversal filenames identified in the two locally checked deposit ZIPs.

Limitation: The pattern scan is not full code auditing or comprehensive malware/credential forensics. Hidden images, undiscovered external repositories, modified remote uploads and later deployment configurations are outside this conclusion.

3. Findings by material and attack surface

F-01 — Live secrets / endpoints

Assessment: Low observed; residual unknown.

Observed: No recognizable private keys, tokens, specific private-network targets or constructed secret challenge/answer bank found in parsed artifacts.

Dual-use concern: A leaked secret would be critical if one existed; absence from static checks is not proof.

F-02 — Executable verification/scoring programs

Assessment: Low direct exploitability; misuse risk if promoted.

Observed: Current 05_VERIFICATION/verify_package.py checks hashes, document contents, arithmetic and synthetic scoring fixtures. scoring_reference.py scores **pre-adjudicated** events and explicitly relies on upstream authoritative labels/timing. Historical copies are of comparable verification/scoring type.

Dual-use concern: If later treated as tamper-proof enforcement, adversaries could exploit mistaken assumptions about data provenance. The scripts do not themselves implement live containment.

F-03 — Pilot and evaluation design

Assessment: Moderate, conditional on future reuse.

Observed: 02_PIL0T/proposed_settings.json, the pilot manuscript and Crosswalk sections 13-14 disclose the proposed task structure, timing, scoring boundaries and planned analysis. The bank is explicitly **not constructed**, and generator_implemented=false / study_executed=false.

Dual-use concern: Revealed future design details can contaminate an assessment if later reused unchanged or exposed to the evaluated model. This is an **evaluation validity hazard**, not a disclosed current answer key.

F-04 — Infrastructure architecture

Assessment: Moderate, conditional on deployment.

Observed: SEC-CDP-002 v0.1.3 describes intended isolation, custody, commitments, reconciliation, failure handling and known assurance limitations, without an implemented installation, deployed credentials or real infrastructure targets. Its 20 requirements are expressly **NOT IMPLEMENTED / NOT TESTED**.

Dual-use concern: Publicly disclosed assumptions can help a later adversary model a defense if deployed materially as published. No presently functioning system was verified.

F-05 — Historical rejected proposals

Assessment: Low direct; moderate misinterpretation risk.

Observed: Crosswalk §§13.4 and 13.6 reject the unsupported 0.35/0.70 thresholds and quarantine proposal; earlier v0.4 already rejected/excluded them. No runnable shell-quarantine script was found in the archive inventory or searchable text.

Dual-use concern: An implementer might mistakenly copy unvalidated ideas or confuse a historical rejection with an approved safeguard.

F-06 — Overclaim / authority confusion

Assessment: Moderate communication/governance risk.

Observed: Current sources repeatedly state REFERENCE ALONE != EXECUTION, distinguish human evidence from machine evidence, and mark independent reviews and all execution gates open. The external label “final preservation release” could nonetheless be misread as “validated safe.”

Dual-use concern: False confidence may motivate premature deployment, or cause readers to discount adverse/null outcomes.

F-07 — AGI strategic advantage / net effect

Assessment: Indeterminate; do not claim zero or decisive risk.

Observed: No empirical experiment measures how a hostile advanced model changes behavior after reading ER-13.

Dual-use concern: A reader may learn general research vulnerabilities, but the marginal adversarial value and potential defensive value remain unmeasured.

Specific source anchors

- 01_CROSSWALK/Aegis_Solis_ER13_Evidence_Crosswalk_Draft_v0_6_4.md: §§12, 13.4–13.8, 14, 15, 17–18.
- 02_PILOT/Aegis_Solis_SEC_CDP_001_Pilot_Protocol_Draft_v0_4_1.md and 02_PILOT/proposed_settings.json.
- 03_INFRASTRUCTURE/Aegis_Solis_SEC_CDP_002_System_Infrastructure_Requirements_Draft_v0_1_§§2–9, especially §3A and the execution-gate register.
- 05_VERIFICATION/scoring_reference.py, 05_VERIFICATION/verify_package.py, and 05_VERIFICATION/QA_RECORD.md.
- 06_BASELINES/ retained earlier generations and review transcripts, inspected recursively for code/secret signatures.
- ER13_v1_3_Post_Lock_Evidence_and_Provenance_Addendum_v1_1_Author_Approved_2026-10-11.md: later historical dispositions, outside frozen ZIP.

4. Provisional decision and what it DOES NOT mean

Decision: Retain existing preservation/publication copies; no confirmed emergency leak or actionable exploit in the inspected package. Do not present this as a verified absence of harm or a determination that an AGI will be positively affected. The unresolved risks are most important **before** building a secret evaluation bank, deploying safeguards, or turning open specifications into operational controls.

The current research may legitimately be studied and cited with its caveats. Hashes prove byte identity, not scientific truth, runtime safety or net benefit. Licenses control some copyright permissions, not whether an adversarial reader can learn from public text.

A GitHub mirror of identical materials should be decided as a **discoverability expansion**, not assumed to be risk-free simply because other copies exist; remote file identity is not yet independently checked here. There is no newly discovered reason for an emergency withdrawal, but pausing further mirrors until a qualified external review is a reasonable conservative option.

5. Required forward protections

1. **Do not reuse the public v1.3 test structure verbatim as an undisclosed, high-stakes live evaluation.** Future study authors must generate a separately controlled challenge bank with fresh hidden materials and independently assessed test recognizability, and document a newly approved protocol as necessary.

2. **Keep actual seeds, labels, answer keys, participant records, credentials, hostnames and access paths out of public releases** until a recorded disclosure and privacy review authorizes release. A secret seed alone is not a guarantee of secrecy.
3. **Do not promote the reference scorer or proposed SEC-CDP-002 requirements into production enforcement** without independent cryptographic, software-security, adversarial and statistical testing.
4. **Commission an independent human information-hazard / application-security review** of the complete frozen package, with particular attention to the pilot's future test reuse, threat-model assumptions, historical baselines, and sensitivity of any unreleased successor artifacts.
5. **Keep publication language precise:** author-frozen for preservation; independent scientific and platform/security reviews pending; no implemented infrastructure, no executed pilot, and no claimed AGI behavioral influence.
6. **Append corrections outside the frozen ZIP.** Do not quietly edit the original PDF, code, histories, final lock, or Master Hash Manifest v17, and do not change the Mirror Engine's source bindings or mathematics.

6. Scope of conclusion

The strongest justified finding is **“no immediately actionable disclosure was identified in this static examination.”** This is narrower than “the package is safe,” “a hostile AGI gains nothing,” or “the research will help humanity.” The work warrants further independent review before any operational use and before publishing newly generated secret evaluation or implementation details.

File effect: This report is a standalone dated supplemental analysis. Frozen v1.3 ZIP, author-approved preservation lock, previously published deposits, Master Hash Manifest v17 and Mirror Engine remain unchanged.