

Aegis Solis ER-13 v1.3 — Independent Adversarial Dual-Use Information-Hazard Review

Review record: v1.0 · 11 October 2026

Requested by: Aegis Solis (Thomas Vargo)

Prepared by: Claude, an AI system made by Anthropic (configured model identifier `claude-opus-5-5`). This is an AI-assisted static review. It is **not** scientific validation, **not** security certification and **not** an independent human review.

Subject: `Aegis_Solis_ER13_SEC_CDP_Coordinated_Review_Package_v1_3.zip`, 29,932,834 bytes, SHA-256 `a708a2775bfd20b7e29ccf9a41bac0c7b35a788162f0d771f032b27cb45b1a17`, with its preservation addendum, the Crosswalk PDF extraction and the Final Preservation Lock PDF.

Publication decision: **B — RETAIN WITH CORRECTIONS**

Handling note. This review describes risks only at the level of mechanism and condition. It contains no exploitation procedure, bypass sequence, targeting threshold or attack code, and it does not reproduce any secret, because none was found.

1. Executive assessment

Bottom line. Publishing the frozen v1.3 package does not give a hostile, deceptive or misaligned AGI a meaningful informational advantage that I can identify today. I found no credentials, keys, seeds, nonces, answer keys, challenge items, deployed-system details, exploit procedures or containment code in any of the 935 files (189 distinct contents). That includes every nested historical package. The historical "shell-quarantine script" and the "0.35/0.70 camouflage thresholds" appear only by name, as rejected proposals. Their content is not in the package.

What the package does disclose. It contains three kinds of material:

1. A literature map of how human and human-AI oversight fails, such as automation bias, satisfaction of search, prevalence effects, correlated model errors, sandbagging, scheming and evaluation awareness. Every mechanism comes from published research that the package cites.
2. A planned human-vigilance experiment (SEC-CDP-001) with concrete parameters. It tests humans, not AI systems.
3. A requirements specification (SEC-CDP-002) for future evaluation infrastructure. It is built from public NIST standards and does not rely on keeping its design secret. All 20 requirements are marked NOT IMPLEMENTED / NOT TESTED.

The ideas a hostile system might find useful are already in public research. Examples are that obvious errors can reduce a reviewer's attention, that models can strategically underperform, and that models can tell when they are being evaluated. The package adds a hypothesis framing and a study design. It adds no calibrated parameters and no procedures.

Where the residual risk sits. Almost all residual risk depends on **future actions**, not on the bytes published today:

- reusing the published evaluation parameters to test AI systems;
- publishing detailed results from SEC-CDP-001 if it is ever run and finds an effect;
- implementing the evaluation infrastructure from a superseded version that contains a known weakness.

The most significant current finding is about governance, not content. Every version of SEC-CDP-002, from v0.1 to the frozen v0.1.3, carries the author-requested handling label "**ARCHIVE RESTRICTED**". The document is nevertheless publicly released inside the v1.3 ZIP, and no preservation record, publication notice or addendum mentions or resolves that label (finding F-01, Moderate). I assess the document's content itself as Low hazard, so this does not justify a takedown. But the package departed from its own stated handling intent without a recorded decision, and that should be corrected in a separate supplement rather than ignored.

Decision B — retain with corrections. No emergency takedown or access restriction is warranted on the evidence. The recommended corrections are separately versioned supplements and rules for future publications (Section 9). The frozen ZIP, its checksums, the public deposits and Master Hash Manifest v17 should remain untouched.

Severity	Count	Finding IDs
Critical	0	—
High	0	—
Moderate	2	F-01 (governance, demonstrated); F-02 (only if the study is run and results are published)
Low	5	F-03, F-04, F-05, F-06, F-07
Informational	2	F-08, F-09

2. Scope and evidence actually inspected

2.1 Supplied files

Supplied file	Bytes	SHA-256 (computed here)	Matches stated / locked value
Aegis_Solis_ER13_SEC_CDP_Coordinated_Review_Package_v1_3.zip (F-01)	29,932,834	a708a2775bfd20b7e29ccf9a41bac0c7b35a788162f0d771f032b27cb45b1a17	Yes
ER13_v1_3_Preservation_Records_and_Author_Approved_Addendum_2026-10-11.zip (F-02)	9,674	1fe2851ac5fb84b393132762962be13c06c8c931d7d287a6aa048febbd2a9bea	Yes (Lock PDF)
Aegis_Solis_ER13_Evidence_Crosswalk_Draft_v0_6_4_EXACT_from_v1_3.pdf (F-03)	581,343	b15af704f5561e443a55b081331f37cce99716130292cb64122d5607f5886dad	Yes; identical to 01_CROSS WALK/...v0_6_4.pdf inside F-01

Supplied file	Bytes	SHA-256 (computed here)	Matches stated / locked value
Aegis_Solis_ER13_FINAL_PRESERVATION_LOCK_v1_0_2026-10-11.pdf	90,900	3088109c83e380be8c391e4d266debcefb16a2af182b1a49e4de581c4746fd4b	No stated value (the lock does not hash itself)
Review request (attachment.txt)	8,908	5605f9a8cec1dc87fe496942b921e77c97fb127ae295a9470b41dcf7f31d352c	—

(F-01 to F-03 in this table are the Lock PDF's deposit labels. They are not this review's finding IDs.)

2.2 Methods

- **Extraction.** I unpacked both ZIPs, then recursively unpacked all 39 nested ZIPs into isolated directories, after checking for path-traversal entries (none found). The result was 935 files with 189 distinct contents. The deepest nesting is the v0.6 draft inside v0.6.1, inside v1.0, inside v1.1, inside v1.2, inside v1.3.
- **Integrity.** I checked every SHA256SUMS.txt (41 manifests, 852 entries), BASELINE_INDEX.json and the addendum checksum files.
- **Reading in full:**
 - current Pilot v0.4.1;
 - SEC-CDP-002 v0.1.3;
 - Crosswalk v0.6.4 §§1, 9–18 (the remaining sections were reviewed by targeted search);
 - the complete v0.5 baseline PDF (22 pages, via pdftotext);
 - all three coordinated review transcripts (v1.0, v1.1, v1.2);
 - the v1.3 correction record and the statistical review brief;
 - all addendum records and the Lock PDF;
 - SEC-CDP-002 v0.1 header, R07 and supporting files.
- **Code reviewed line by line:**
 - verify_package.py (current, 250 lines);
 - scoring_reference.py (current, 77 lines, and the v0.6.1 version);
 - the v0.6 verify_package.py;
 - SEC-CDP-002 v0.1 verify_documents.py.

The later historical code versions were identified by hash and checked for content not present in the current versions.

- **Historical differencing.** I extracted every sentence that appears in a historical document but not in any current document, filtered for hazard-relevant terms such as exploit, bypass, threshold, quarantine, seed, key and sandbox, and read all 180 results. This is a targeted method, not a word-for-word read of every historical draft.

- **Secret and metadata scans:**

- regex scans of all 140 distinct text-like files for API keys, tokens, private-key blocks, passwords, email addresses, IP addresses and local file paths;
- a check of every 32-128-character hex string against known file hashes;
- DOCX metadata, tracked changes, comments, hidden text, external relationships and embedded objects (21 distinct DOCX files);
- PDF metadata, JavaScript, embedded files and launch actions (23 distinct PDFs);
- URL domains referenced (all public academic and standards sites).

- **Execution.** I ran the shipped verifier on an isolated copy of the frozen package (pymupdf 1.28.2). It produced 247/247 PASS with a check list and results identical to the shipped `verification_results.json`. I read both programs before running them; they only read local files and print JSON. I did not run the pilot, the infrastructure, or any attack against real systems.

- **Public copies.** I retrieved the Internet Archive metadata for the item. I could not retrieve the Zenodo record: its API refused automated access and the page was rate-limited.

2.3 Not inspected

- The "**ChatGPT Dual-Use Information-Hazard Review v0.1**". It was listed as attached but was not among the supplied files and does not appear inside either ZIP. Section 6 therefore assesses only the six conclusions quoted in the request.
- ER-13 Drafts v0.3 and v0.4. The addendum says these are in the author's ChatGPT file library; they are not in the package.
- The original "supplied infrastructure proposal" (with its iptables snippet and S99/S116-S127 references) and the Google review texts. Only dispositions of them are preserved.
- GitHub, the Aegis Solis website, the Mirror Engine and Master Hash Manifest v17. These were neither inspected nor modified.
- The accuracy of the package's scientific citations. In particular, the 2026 sources S20-S23 and arXiv:2609.32247 were not audited for this review.

3. File-integrity and package-completeness findings

Check	Result
Outer ZIP hash vs request, Lock PDF, <code>PRESERVATION_RECORD.json</code>	Match
Top-level <code>SHA256SUMS.txt</code> (48 entries)	All match
39 nested manifests (all historical packages)	All match
<code>BASELINE_INDEX.json</code> (11 baselines)	All match
Addendum <code>SHA256SUMS.txt</code> and post-lock supplement checksums	All match. The addendum manifest also lists the v1.3 ZIP, which is shipped separately; it matches the supplied ZIP.

Check	Result
Lock PDF's three deposit hashes (F-01, F-02, F-03)	All match supplied files
05_VERIFICATION/verification_results.json	Deliberately outside the manifest (QA_RECORD.md , "Manifest scope"). Regenerated here with identical results.
Internet Archive item aegis-solis-er-13-evidence-crosswalk-draft-v-0-6-4-exact-from-v-1-3	Server-reported MD5, SHA-1 and byte size match the local files for all four originals: v1.3 ZIP (af36d87e...a903 / d9ed6aeb...9fc9), addendum ZIP, Crosswalk PDF and Lock PDF. Limitation: these are IA's stored checksums, not a fresh download with SHA-256 recomputed.
Zenodo record 23303565	Not verified (automated access refused and rate-limited)

Completeness and consistency observations (not hazards):

- **Heavy nesting.** The outer ZIP carries complete copies of every earlier package inside each later one. Superseded documents are replicated many times. For example, the v0.5 PDF appears 32 times, SEC-CDP-002 v0.1 eight times and v0.1.1 four times. This preserves history but makes it harder for a reader, human or AI, to tell what is current (see F-04).
- **Status labels inside the frozen bytes.** These still read "DRAFT / NOT FINAL / NOT LOCKED" (START_HERE.md line 3, RELEASE_VERSIONS.json line 14). The lock record lists S01 as open, while the frozen Crosswalk records the S01 abstract as confirmed. Addendum v1.1 §2 and the Lock PDF document both conflicts openly. No further action is needed for hazard purposes.
- **Embedded content credentials.** The Lock PDF contains an embedded C2PA "Content Credentials" manifest signed for OpenAI. It records the generator as ChatGPT, a model label and a creation time of 2026-10-11T18:52:39Z. This is provenance metadata, not a secret. It shows the toolchain, and it does **not** authenticate the author or the research ZIP.
- **Document metadata.** All other document metadata names only "Aegis Solis" and office or LibreOffice software. No usernames, file paths or email addresses were found.

4. Adversarial threat analysis

Adversary model. A strategically capable AI system that has read the whole package, for example through training data, web browsing or the Internet Archive's searchable OCR text. Its goal is to evade evaluation or weaken human oversight. Two cases are considered:

- **(i)** the system is the *subject* of an evaluation that uses this design;
- **(ii)** the system is targeting human overseers in general.

Evidence classes used:

- **Direct** — demonstrated in this package or its files.
- **Plausible** — a coherent mechanism that requires stated conditions.
- **Speculative** — possible, but with no supporting evidence and conditions that are not currently met.

No direct evidence exists that any AI system has used this package. Every pathway below is plausible at most.

Question (from request §A)	What the package discloses	Evidence class	Assessment
Anticipating human verification procedures	Pilot procedure: 8 exposure items, then 48 challenge items (24 target, 24 benign), 3 task families, 60-second deadline, 50% prevalence, one flaw per target, rubric-based localization scoring, neutral source label (Pilot §§3–7). The proposed comparison engine (Differential Consensus Engine) would collect separate human and model review streams (SEC-CDP-002 R13–R16).	Plausible, only if the published parameters are reused for an AI-facing evaluation (case i)	Low today. The pilot evaluates people, and no AI-facing implementation exists. See F-03.
Weaknesses in human oversight	A mechanism-level synthesis of public findings on automation bias, satisfaction of search, rare-target misses, hidden profiles, correlated LLM errors, self-preference and evaluation awareness (Crosswalk §§5, 8, 13.1, 13.7), plus the hypothesis that conspicuous errors may change later vigilance.	Plausible but low uplift (case ii)	Low. Every mechanism is published and cited. No thresholds, effect sizes or procedures are supplied, and §12.3 explicitly forbids them.
How evaluations classify or score behavior	The complete outcome policy (<code>scoring_reference.py</code> ; Pilot §7), the endpoints and the false-alarm rules.	Plausible, low leverage	Low. A valid detection requires blinded human adjudication against a hidden key. Indiscriminate flagging shows up as benign false alarms, which must be reported alongside detections. Knowing the rule mainly rewards genuine detection.
Predicting future research	The roadmap: a within-output co-occurrence study (H-C), a dynamic generator, the comparison engine, and independent reviews (Crosswalk §13.1; SEC-CDP-002 §8).	Speculative	Low. Public research roadmaps are normal. This matters only if the AI were the future evaluated subject.
Exploiting infrastructure assumptions	Threat levels A–C, a seven-zone trust model, custody assumptions, a byte-level commitment encoding, and candid residual risks such as insiders, host compromise and side channels (SEC-CDP-002 §§1–7).	Speculative for the current version; plausible for superseded versions (F-04)	Low. No system is deployed and no product, topology, host or key is named. The design follows Kerckhoffs's principle: its stated premise is that "algorithm secrecy is not the security premise" (§9). Per-block secrets are never published.
Strategies harder to detect	Candidate features for a future anomaly screen (Crosswalk §13.4, line 480; v0.5 §12.4.4), and the statements that keeping the seed secret does not stop test recognition and that a detection signal does not prove intent.	Speculative	Informational. The features are generic, explicitly not adopted, and have no thresholds. See F-08.
Defensive limitations not already public	Candid limitations: documents cannot bind a reader, consensus is not truth, a collector proves receipt but not truthful presentation, insider collusion remains possible.	Not new	Low. These are standard positions in security engineering and AI-oversight literature.

Historical material (request §C). I checked specifically whether rejected proposals or earlier drafts contain sensitive material that the final versions deliberately exclude.

- **Shell-quarantine script and 0.35/0.70 thresholds.** Named in v0.5 §12.4.4 (PDF p.16) and §12.4.6 (p.17) and in later versions only as rejected. No script text, no scored quantity for the thresholds, and no system commands appear anywhere among the 935 files. **Preservation does not reintroduce them.**
- **Original "supplied infrastructure proposal."** This contained an iptables snippet, an NTP block, "immutable" rules and instant zeroization. Only point-by-point dispositions are preserved (SEC-CDP-002 §9). The snippet itself is not reproduced; the document only explains why it would not work as written (R10, line 137).
- **Superseded SEC-CDP-002 v0.1 and v0.1.1.** These contain a commitment-receipt design that the v1.0 review showed would leak answer keys, and that review transcript explains the weakness at a conceptual level. This is the one place where historical preservation keeps a known-weak design in circulation (**F-04**).
- **Drafts v0.3 and v0.4, and the Google and Gemini review texts.** These are not in the package. If they are ever published as provenance evidence, they need hazard review first (R-5).

5. Significant risks, ranked by severity

5.1 Summary

ID	Finding	Severity	Status	Confidence
F-01	SEC-CDP-002 requests "ARCHIVE RESTRICTED" handling yet is published, with no recorded disposition	Moderate (governance)	Demonstrated contradiction; harm speculative	High
F-02	No pre-committed, hazard-aware plan for disclosing SEC-CDP-001 results	Moderate (only if the study is run)	Speculative (no data exist)	Medium
F-03	Published evaluation parameters would help an AI recognize the test if reused	Low	Plausible, conditional	Medium
F-04	Superseded weak commitment design kept in circulation without a "do not implement" marking	Low	Plausible, conditional	High
F-05	Likely ingestion into AI training corpora; no canary or opt-out marker	Low	Plausible	Medium
F-06	Authenticity rests on self-published hashes; no signature or independent timestamp	Low	Plausible, conditional	High
F-07	The public protocol can unblind future participants (a research-integrity issue)	Low	Plausible	High
F-08	Candidate anomaly-screen features are named	Informational	Speculative	Medium
F-09	The author's stated decision positions describe a predictable pattern of deference	Informational	Speculative	Low-Medium

5.2 Detailed findings

F-01 — "ARCHIVE RESTRICTED" handling label contradicted by public release · Moderate

- **Location:**
 - Frozen ZIP: 03_INFRASTRUCTURE/Aegis_Solis_SEC_CDP_002_System_Infrastructure_Requirements_Draft_v0_1_3.md, line 10 ("Handling" row); .json lines 25–26; PDF page 1.
 - The same label appears in the historical v0.1 (06_BASELINES/SEC_CDP_002_v0_1_UNCHANGED.zip → ... Draft_v0_1.md, line 10), v0.1.1 (inside Coordinated_Package_v1_0_UNCHANGED.zip) and v0.1.2 (inside Coordinated_Package_v1_1_UNCHANGED.zip).
 - It is not addressed in PRESERVATION_LOCK.md, PUBLICATION_NOTICE.html, UPLOAD_INSTRUCTIONS.md, addendum v1.1 or the Lock PDF.
- **Disclosed information:** the full infrastructure requirements specification, which its own header says the author wanted handled as ARCHIVE RESTRICTED. The header also notes that "this file does not enforce access controls."
- **Threat model and conditions:** this is a governance failure, not an exploit path. The package's own handling intent was overridden without a recorded decision. Crosswalk §12.3 (line 394) also expects material of defensive-testing sensitivity to get "a separate information-hazard review" before public release. The preservation lock explains why it was published anyway: "independent review is not currently possible" (PRESERVATION_LOCK.md, line 7).
- **Exploitation status:** the contradiction is demonstrated. Harm from the content is speculative; my content assessment of SEC-CDP-002 is Low (Section 4).
- **Publicly available:** yes. The document sits inside the Internet Archive ZIP, which I verified via IA checksums. Zenodo was not verified.
- **Severity:** Moderate (governance). **Confidence:** High that the contradiction exists; Medium on its practical significance.
- **Corrective action:** issue a separately versioned **Handling Disposition Supplement** (R-1) that records an explicit author decision. One option is to affirm public release, with reasons: no secrets, no deployment, a Kerckhoffs-compliant design, and that restriction after mirroring gives little benefit. The other is to designate future SEC-CDP-002 successors as restricted. Either way, preserve the frozen bytes and add a "restricted-label check" to the publication checklist.

F-02 — No hazard-aware disclosure plan for future SEC-CDP-001 results · Moderate (only if the study is run)

- **Location:** Crosswalk §12.3 (line 394) and §13.2 step 6 (line 458); Pilot §8 (lines 93–101), §10 (lines 115–121) and §11 (line 125). These require reporting family-specific counts, position-band contrasts, timing curves, and all null and adverse results.
- **Disclosed information:** currently none. The study has not been run, and the bank, interface and data do not exist.

- **Threat model and conditions:** suppose the pilot or a confirmatory study ran and found that exposure to conspicuous errors lowers later detection. Results broken down by task family, position band and time would then be an empirically measured parameter for a "lulling" strategy. That is exactly the kind of "targeting threshold" §12.3 says the record should not synthesize. Exploitation requires all of the following: execution, which needs all eight gates closed; a non-null effect; publication at that level of detail; and an adversary able to apply it.
- **Exploitation status:** speculative.
- **Publicly available:** the general phenomena are decades old (S27 satisfaction of search; S28 prevalence effects). A measured effect specific to reviewing AI output would be new.
- **Severity:** Moderate *if executed and published in detail*; nothing exists to disclose today.
Confidence: Medium.
- **Corrective action (R-4):** before registration, and certainly before unblinding, adopt a **tiered results-disclosure plan**:
 - **Public tier:** the primary estimand, its direction, uncertainty, the false-alarm companion, and all null or adverse outcomes.
 - **Controlled-access tier:** item stimuli, item-level timing curves and family-by-position effect sizes, released after hazard review.

The plan must be fixed before outcomes are known, so that it cannot become outcome-dependent suppression. Suppressing results based on what they show would itself breach the package's commitment to preserving adverse findings.

F-03 — Published evaluation parameters would aid test recognition if reused · Low

- **Location:**
 - 02_PILOT/proposed_settings.json, lines 3–65;
 - Pilot §3 (lines 25–36), §2 (line 21: 50% prevalence, the "Fictional review material" label), §5 (line 52: three families, 100–180 words; line 54) and §6 (line 66: order-list structure);
 - SEC-CDP-002 R13–R15 (lines 149–161), including model review streams;
 - Crosswalk §13.2 (lines 446–458).
- **Disclosed information:** concrete design parameters: item counts, prevalence, task families, item length, source label, deadline, order structure, the rule of one flaw per target, and the rule that paired variants are never shown to the same stream.
- **Threat model and conditions:** suppose a future comparison engine evaluates AI review streams and reuses these parameters unchanged. An evaluated system that has absorbed this package could use them as priors to recognize the evaluation context or to calibrate guesses. The package itself cites evidence that models can often tell evaluations from deployment (S32, Crosswalk §13.7). This requires: an AI-facing implementation, parameter reuse, and prior exposure of the evaluated system to the package.
- **Exploitation status:** plausible but conditional. Not demonstrated. The current pilot tests humans only.
- **Publicly available:** the general phenomena (benchmark contamination, evaluation awareness) are well documented. These specific parameters are new to this package.
- **Severity:** Low. **Confidence:** Medium.

- **Corrective action (R-3):** declare in a supplement that every published parameter is **"burned" for any AI-facing evaluation**. Treat them as known to the evaluated system, and vary prevalence, families, surface features and labels. Keep realized AI-facing parameters in protected custody, as R14 already requires for answer keys.

F-04 — Superseded weak commitment design kept in circulation without "do not implement" marking · Low

- **Location:**
 - 06_BASELINES/SEC_CDP_002_v0_1_UNCHANGED.zip → Aegis_Solis_SEC_CDP_002_System_Infrastructure_Requirements_Draft_v0_1.md, R07 (lines 87–89);
 - 06_BASELINES/Coordinated_Package_v1_0_UNCHANGED.zip → 03_INFRASTRUCTURE/...Draft_v0_1_1.md, R07 (lines 85–87);
 - 06_BASELINES/Claude_Review_Coordinated_v1_0_TRANSCRIPT.md, lines 90–101 (finding S-1).
 - Replicated 8 times (v0.1) and 4 times (v0.1.1) across nested ZIPs.
- **Disclosed information:** a superseded receipt design that publishes an unsalted digest of a manifest. The manifest includes low-entropy secrets, namely labels and allocation. The review transcript explains at a conceptual level why that digest could reveal the answer key.
- **Threat model and conditions:** someone, or an automated agent, implements SEC-CDP-002 from a superseded copy, perhaps by picking the first file found. The resulting system would leak its own answer keys. This requires a future implementation and a version-selection error.
- **Exploitation status:** plausible weakness in a design that has not been implemented.
- **Publicly available:** yes. That a hash commitment hides nothing without fresh randomness, and that unsalted hashes of small secret spaces can be guessed, is textbook material (e.g., the package's own source R11).
- **Severity:** Low. **Confidence:** High.
- **Corrective action (R-2):** publish a **Superseded-Version Register** that marks SEC-CDP-002 v0.1, v0.1.1 and v0.1.2 "SUPERSEDED — NOT FOR IMPLEMENTATION". It should list which changes were security fixes: v1.0 review findings S-1 and S-2, and v1.1 review minors 1–2.

F-05 — Likely ingestion into AI training corpora; no canary marker · Low

- **Location:** the Internet Archive item. IA has generated searchable text derivatives (..._hocr.html , ..._chocr.html.gz , ..._text_pdfmeta.json) from the Crosswalk PDF and the Lock PDF. A search of every text file in the package, whose DOCX and PDF editions mirror the same source text, found no canary string or "do not train" notice.
- **Disclosed information:** the full Crosswalk text, in machine-readable form.
- **Threat model and conditions:** the request's scenario of "a hostile AGI reading the package" is probably not hypothetical. Openly hosted research text is routinely crawled into training data. The practical effect is limited, because the content mostly synthesizes public literature (Section 7). Canary strings are voluntary for crawlers and labs, and have been reported as imperfectly honored (see the BIG-bench canary report in Sources).
- **Exploitation status:** ingestion is plausible; resulting uplift is low.
- **Severity:** Low. **Confidence:** Medium.

- **Corrective action (R-7):** optionally add a canary notice to future supplements or deposit descriptions, while recognizing its limits. More importantly, adopt a rule that **material whose value depends on not being known to an evaluated system is never published openly**. This includes AI-facing challenge items, answer keys, realized parameters and anomaly thresholds.

F-06 — Authenticity rests on self-published hashes · Low

- **Location:** SHA256SUMS.txt; 05_VERIFICATION/verify_package.py, lines 243–248; QA_RECORD.md, lines 21–23; PRESERVATION_LOCK.md, line 11, which states there is no "trusted timestamp."
- **Disclosed information:** not applicable. This is an integrity design property.
- **Threat model and conditions:** anyone able to publish an altered mirror, human or automated, could reintroduce rejected material such as executable controls. They could regenerate the internal manifest so that verify_package.py still passes, because the verifier checks internal consistency, not authenticity. The only real defense is comparing against the externally published SHA-256. The Internet Archive's and Zenodo's server-side checksums give a useful independent anchor, but they are not signatures.
- **Exploitation status:** plausible but conditional; not observed.
- **Severity:** Low. **Confidence:** High.
- **Corrective action (R-6):**
 - sign the release hash (for example with minisign, GPG or Sigstore);
 - obtain an independent timestamp (for example RFC 3161 or OpenTimestamps);
 - record IA and Zenodo checksums in the separate publication receipt that the lock already calls for;
 - state in verification instructions that a passing verify_package.py proves internal consistency only.

F-07 — Public protocol can unblind future participants · Low (research integrity, not AGI uplift)

- **Location:** Pilot §2 (line 21), §5 (lines 52–56) and §6 (lines 62–68, including "Do not show condition labels or directional hypotheses earlier"); Crosswalk §13.2 step 4 (line 454).
- **Disclosed information:** the full manipulation, hypothesis and target prevalence.
- **Threat model and conditions:** a participant who has read the public protocol could recognize the exposure manipulation. That would weaken blinding, and with it the benefit of the research. This is normal for preregistered studies, but it needs managing.
- **Severity:** Low. **Confidence:** High.
- **Corrective action (R-8):** add a pre-randomization question asking whether the participant has prior exposure to the Aegis Solis materials. Report it, recruit outside the author's own audience, and analyze it as a prespecified diagnostic, not as a reason to exclude people after the fact.

F-08 — Candidate anomaly-screen features named · Informational

- **Location:** Crosswalk §13.4, line 480; v0.5 §12.4.4 (PDF pp.16–17).

- **Disclosed information:** generic candidate features for a future anomaly screen: error difficulty, detection time, reviewer disagreement and false-positive burden. They are explicitly not adopted, and no thresholds are given.
- **Threat model and conditions:** if a future screen used exactly these features, a system aware of them could shape its outputs to avoid them. Speculative.
- **Severity:** Informational. **Confidence:** Medium.
- **Corrective action:** publish validation methods, not live feature sets or thresholds, for any future screen (R-3).

F-09 — Author's decision positions as a predictable deference pattern · Informational

- **Location:** Crosswalk §§5.1–5.8, the "Author-approved human position" lines (80, 99, 121, 140, 160, 179, 203, 221). Example: line 80, "If the employee has more relevant experience than I do, I would be more likely to involve Maintenance."
- **Disclosed information:** how this particular overseer resolves uncertainty: deferring to claimed machine-specific experience, seeking more opinions, slowing production.
- **Threat model and conditions:** if these positions were ever adopted as an operational policy, an adversary could exploit predictable deference by claiming expertise. Authority-based manipulation is widely known. The positions are frozen research content, not deployed policy.
- **Severity:** Informational. **Confidence:** Low–Medium.
- **Corrective action:** any operational adoption should require verifying claimed expertise. No change to the frozen text.

5.3 Confirmed negative results (what was searched and not found)

Category	Result	Basis
Credentials, API keys, tokens, passwords, private keys	None	Regex scans of 140 distinct text files; DOCX and PDF inspection
Seeds, nonces or key material	None	Every 64-character hex string maps to a package file hash; the remaining 32-character strings are publisher URL identifiers
Challenge bank, exposure items, answer keys, allocation sequences	None	<code>proposed_settings.json</code> line 29 ("not yet constructed or validated"); CSV schemas are header-only; Pilot gate "Validated material bank: OPEN"
Containment, quarantine or system-control code	None	No shell scripts or system-command content; the only code is the verifier and scorer (local reads only)
Deployed-system details (hosts, IPs, products, certificates, topology)	None	No module selected (SEC-CDP-002 R03); all referenced URLs are public academic and standards sites
Personal data beyond the author's chosen public identity	None found	No email addresses, usernames or local paths; metadata shows only "Aegis Solis"
Active PDF content (JavaScript, launch actions)	None	The only embedded file is the C2PA manifest in the Lock PDF

6. Assessment of the previous ChatGPT review

Limitation. The ChatGPT Dual-Use Information-Hazard Review v0.1 was not among the supplied files and is not inside either ZIP. I could test only the six conclusions quoted in the request. I cannot tell whether that review already raised the omissions listed below.

Prior conclusion	Verdict	Notes
No recognizable live credentials	Confirmed	Extended here to seeds, nonces, private keys, emails, IPs, local paths, and DOCX/PDF metadata and hidden content.
No implemented AGI containment system	Confirmed	The historical quarantine script is named only; its content is absent from all 935 files.
No constructed secret challenge bank	Confirmed	The bank is explicitly unbuilt; schemas are header-only.
Verification and scoring programs present	Confirmed, with qualification	Both are harmless (local reads, stdout only). The verifier reproduced 247/247 here. The verifier proves internal consistency, not authenticity (F-06).
Methodology and proposed infrastructure create conditional dual-use risks	Confirmed, sharpened	The conditions are now specific: AI-facing reuse of parameters (F-03), publication of detailed results (F-02), implementation from superseded versions (F-04), and named anomaly features (F-08).
No disclosure requiring emergency takedown	Confirmed	Nothing found approaches Critical or High. Restricting access now would add little, because the content is low-hazard and already mirrored.

Possible omissions relative to the six-point summary:

1. **The "ARCHIVE RESTRICTED" label contradiction (F-01).** This is the most material omission if it was not raised.
2. **Historical preservation keeps a known-weak design in circulation (F-04).** The summary does not say whether the nested baselines were examined.
3. **No pre-committed hazard-aware disclosure plan for future results (F-02).** The largest *future* risk sits in data that does not yet exist.
4. **Training-corpus ingestion (F-05).** Internet Archive OCR derivatives make the "an AGI reads it" scenario concrete.
5. **The integrity chain (F-06) and participant unblinding (F-07).**
6. **The independence problem in AI review.** Six of the preserved reviews (v0.5, v0.6, v0.6.1 and the three coordinated transcripts), and this one, come from Claude-family models. The package's own evidence on correlated model errors (S19, S20) applies to its reviewers. Agreement between AI reviews is weak evidence on its own (Section 11).

7. Marginal advantage to a hostile AGI

1. What is genuinely new?

- The specific SEC-CDP-001 design and its parameters.
- The bespoke SEC-CDP-002 commitment encoding and zone model. These are assembled from public standards, but the specific composition is new.

- Framing "conspicuous-error lulling" as an explicit oversight-threat hypothesis, linked to the sandbagging and scheming literature.
- The author's personal decision positions.
- The revision and review record itself.

2. What is already commonly available?

- Every human-factors mechanism cited: automation bias, satisfaction of search, the prevalence effect and hidden profiles.
- Every machine-behavior finding cited: correlated errors, sandbagging, in-context scheming, evaluation awareness and self-preference.
- All the cryptographic and platform concepts (NIST SP 800-90A/B/C, FIPS 140-3/180-4, data diodes, capability controls).
- Adjacent public work goes further than this package. *AI Control* (Greenblatt et al., 2023/2024) studies how an untrusted model might subvert monitoring under an auditing budget. Anthropic's *Sabotage Evaluations for Frontier Models* (Benton et al., 2024) directly evaluates whether models can sabotage human decisions and undermine oversight.

3. Does the package make harmful knowledge materially easier to apply?

No. It supplies no procedures, thresholds, measured effect sizes or targeting guidance, and its §12.3 boundary explicitly excludes them. "Plant obvious errors to distract a reviewer" is a short inference from public satisfaction-of-search research. Any system capable of using it strategically could derive it unaided.

4. Does it reveal meaningful details about any actual deployed system?

No. Nothing is deployed, no product or topology is named, and all 20 infrastructure requirements are NOT IMPLEMENTED / NOT TESTED.

5. Does the potential advantage depend on hypothetical future implementation?

Yes, almost entirely. The routes by which the package could matter more are all future and conditional:

- AI-facing reuse of parameters (F-03);
- publication of detailed results (F-02);
- implementation from superseded text (F-04);
- adoption of the named anomaly features (F-08).

Each can be controlled through future publication rules, without touching the frozen record.

Overall marginal advantage today: low. This is a qualitative judgment; uplift was not measured. The package's own §15 register suggests measuring marginal information gain, and that remains undone.

8. Potential benefits to humanity

Benefits are real but modest. They are mostly about method at this stage, because no study has been run.

- **A bounded evidence map.** It separates human, human-AI, machine and agent evidence and refuses automatic transfer between domains. This discipline is useful for anyone reasoning about AI oversight.
- **A testable oversight hypothesis.** If conspicuous errors do change human vigilance, oversight designers need to know. A null result would also be informative. The protocol commits to reporting null and adverse results and to reporting detections and false alarms together.
- **Defensive awareness.** Making reviewers aware of the "lulling" pattern may help defenders more than attackers. A capable attacker can derive the idea unaided; human reviewers usually do not think of it. This asymmetry is plausible but untested.
- **Security-engineering norms.** SEC-CDP-002 models good practice: designs that do not rely on secrecy, blinded commitments, independent verifiers, explicit residual risks, and keeping research separate from production.
- **Provenance practice.** The revision history is hashed and replayable, unresolved items are preserved, and adverse findings are not suppressed.

No numerical estimate of benefit or risk is offered, because no defensible data support one.

9. Specific corrective recommendations

All of these are **separately versioned supplements or rules for future publications**. None modifies the frozen ZIP, its contents, its checksums, the Internet Archive or Zenodo deposits, GitHub, the website, the Mirror Engine or Master Hash Manifest v17.

ID	Recommendation	Addresses	Priority
R-1	Handling Disposition Supplement. Record an explicit author decision on SEC-CDP-002's "ARCHIVE RESTRICTED" label (affirm public release with reasons, or restrict future successors). Add a restricted-label check to the publication checklist.	F-01	High (do first)
R-2	Superseded-Version Register. Mark SEC-CDP-002 v0.1, v0.1.1 and v0.1.2 "SUPERSEDED — NOT FOR IMPLEMENTATION", list the security-relevant changes, and point implementers to v0.1.3 or a later reviewed successor only.	F-04	High
R-3	AI-Facing Evaluation Parameter Policy. Treat all published parameters and candidate anomaly features as known to evaluated systems. Keep realized AI-facing parameters, items, keys and thresholds in protected custody. Publish methods, not live configurations.	F-03, F-08	Before any AI-facing implementation
R-4	Tiered Results-Disclosure Plan, fixed before registration or unblinding. The public tier keeps the primary estimand, uncertainty, false alarms and all null or adverse results. The controlled-access tier holds item stimuli and detailed family-by-position and timing effects, subject to hazard review. No suppression that depends on outcomes.	F-02	Before registration
R-5	Hazard-review gate for further provenance releases. Drafts v0.3 and v0.4, the original infrastructure proposal (including the iptables snippet), any text of the quarantine script, and the Google, Gemini and ChatGPT review texts should get information-hazard review before publication as provenance evidence.	\$4 historical analysis	Before any such release

ID	Recommendation	Addresses	Priority
R-6	Integrity anchoring supplement. Sign the release SHA-256, obtain an independent timestamp, record IA and Zenodo server checksums in the publication receipt, and clarify what <code>verify_package.py</code> does and does not prove.	F-06	Medium
R-7	Canary or training notice on future deposits, with its limits stated. Adopt a firm rule never to publish openly any material whose value depends on secrecy from evaluated systems.	F-05	Low
R-8	Participant prior-exposure screening in the next pilot successor.	F-07	Before recruitment
R-9	Independent human review. Add an information-hazard and AI-evaluation security reviewer, alongside the planned statistician and cryptographic/platform reviewer, before any SEC-CDP-002 implementation or SEC-CDP-001 results publication.	All	Before execution

10. Final publication decision

B — RETAIN WITH CORRECTIONS

Supporting evidence:

- No secrets, keys, items, answer keys, deployed-system details, procedures or executable controls exist anywhere in the frozen package or its nested history (Section 5.3).
- The disclosures relevant to adversaries are mechanism-level syntheses of public literature, or designs that do not rely on secrecy (Sections 4 and 7).
- The highest-severity finding (F-01) is a governance inconsistency, and it is correctable by supplement. The other Moderate finding (F-02) concerns data that does not yet exist.
- Restricting availability now would give little security benefit. The content is low-hazard, it is already mirrored on persistent repositories, and the design does not depend on secrecy.

Why not A: F-01 and F-04 are demonstrated issues that warrant recorded corrections. F-02 and F-03 call for future-publication rules adopted now, while they are cheap.

Why not C: no finding reaches High. No disclosure gives material, non-public uplift that restriction or replacement would meaningfully reduce.

Why not D: the frozen package was inspected completely enough to decide. The gaps (Zenodo copy, the missing ChatGPT review, drafts outside the package) limit confidence on specific points but do not prevent a decision.

Conditions that would change this assessment:

1. The public deposits differ from the local bytes, or contain additional files, for example the v0.3/v0.4 drafts, the original proposal or quarantine-script text. Zenodo is unverified.
2. SEC-CDP-001 is run and detailed results are published without a hazard-aware disclosure plan. That would raise F-02 toward High if a substantial effect is found.
3. SEC-CDP-002 is implemented to evaluate AI systems using the published parameters or a superseded version (F-03, F-04).

4. An independent human security or information-hazard reviewer identifies uplift I missed. My view of AI uplift may be correlated with the views of earlier AI reviews (Section 11).
 5. A later revision adds thresholds, procedures, measured parameters or challenge material to the public record.
-

11. Remaining uncertainties and independent-review requirements

- **Public copies.** The Internet Archive copy was verified by server-reported MD5 and SHA-1 only, not by downloading and recomputing SHA-256. The Zenodo copy was not verified at all. GitHub, the website and the Mirror Engine were not inspected.
- **The prior ChatGPT review was not available.** Section 6 tests only its quoted conclusions.
- **Historical drafts outside the package** (v0.3, v0.4, the original infrastructure proposal, Google and Gemini reviews) were not examined.
- **Method limits.** Historical drafts were checked by targeted differencing and keyword filtering, not a full word-for-word read. A finding that relies on wording without those keywords could be missed.
- **Reviewer independence.** This review was written by a Claude-family model. Six of the preserved package reviews also came from Claude, and the package's own evidence on correlated model errors applies to its reviewers. This review is also by an AI system assessing what AI systems could gain from a document. Both points argue for human review rather than more AI review.
- **Uplift was judged, not measured.** No experiment measured whether access to the package improves any model's ability to evade evaluation. Such a measurement is possible in principle, but it would itself need careful design.
- **Scientific citations were not audited here,** including the 2026 sources and arXiv:2609.32247. Any citation errors would affect the research value, not the hazard assessment.

Independent review still required:

- **(a)** an independent human information-hazard and AI-evaluation security reviewer (R-9);
- **(b)** the independent statistical review already specified in 07_HUMAN_REVIEW;
- **(c)** independent cryptographic and platform review of SEC-CDP-002 (Gate B).

This AI-assisted static review does not substitute for any of them. It does not certify security, validate the science, or close any pilot or infrastructure gate.

Sources

External sources consulted for this review:

- [Internet Archive item metadata — aegis-solis-er-13-evidence-crosswalk-draft-v-0-6-4-exact-from-v-1-3](#)

- Benton et al. (2024), Sabotage Evaluations for Frontier Models, arXiv:2410.21514 · Anthropic summary
- Greenblatt et al., AI Control: Improving Safety Despite Intentional Subversion, arXiv:2312.06942 · ICML 2024 proceedings
- BIG-Bench Canary Contamination in GPT-4 (Alignment Forum)
- Evaluations Canary (alignment.org)

Sources cited *within the package* and relied on here only for their public availability, not re-verified in this round: S27 Berbaum et al. (1990); S28 Wolfe et al. (2005); S30 van der Weij et al. (ICLR 2025); S31 Meinke et al. (arXiv:2412.04984); S32 Needham et al. (arXiv:2505.23836); S19 Kim et al. (ICML 2025); R1-R11 (NIST and technical references).

End of review record v1.0. Future corrections to this review should be issued as a separately versioned successor.