

ER13 Draft v0.6 Change and verification record

Aegis Solis | 10 October 2026 | NOT FINAL / NOT LOCKED

This record documents the targeted successor from the supplied v0.5 PDF and Claude critique. It separates completed manuscript corrections from historical attribution gaps and unexecuted empirical work.

1. Input identity

The supplied v0.5 PDF SHA-256 is b9d6e4d18bd9aa73144f202ec6f5faf16bc92e10428691bc98223625bcd29ad9, matching the prior response. Its exact bytes and the supplied Claude review are retained in 04_BASELINE. The earlier v0.5 ZIP and editable DOCX were not supplied and their historical hashes were not verified.

The successor was reconstructed from all 22 PDF pages. Eight canonical IC questions and eight author-approved positions are retained verbatim after whitespace normalization. Original source identifiers S01-S26 are retained; S27-S34 are added. Table layout is reflowed into readable fields, and source labels are restored from the source register to remove PDF continuation artifacts.

2. Review dispositions

Finding	Disposition	Change
B1	ACCEPT	S14 finding corrected; DVD/private-preference context retained; section 5.1 support sentence revised.
M1	ACCEPT	H-S and H-C separated in 13.1. Current experiment selects H-S; H-C remains explicitly untested. S27-S29 provide competing human mechanisms.
M2a	ACCEPT	One primary marginal risk difference Delta across 13.3 and 14.1-14.4. Conditional coefficients secondary; integration and simulation target Delta.
M2b	ACCEPT	Joint item-type by condition model; mandatory benign false-alarm contrast; secondary signal-detection summaries with assumptions.
M2c	ACCEPT	Undetected items administratively censored for timing; restricted mean/curves replace detected-only median contrast.
M2d	ACCEPT	Order_ij; recurring items across arms; crossed evaluator and item effects, including item exposure heterogeneity.
M3	ACCEPT WITH BOUNDARIES	S30-S34 added. Strategic underperformance and evaluation awareness separated from conspicuous-error manipulation, natural deployment, and general AGI behavior.
Minor 1	ACCEPT	A added to IC-EPI-002 category.
Minor 2	ACCEPT	S24-S26 Domain/Status/Access fields standardized and explicit roles cited in 14.7.
Minor 3	PARTLY ACCEPT	History block added; sections 6/7/11 retained as historical records; five response topics separated from the inherited defect count.
Minor 4	PARTLY RESOLVED	13.4 records earliest supplied location and unadopted status. Original proposer and first version remain unverified; no attribution invented.

Finding	Disposition	Change
Minor 5	ACCEPT	13.2 permits fixed library or preregistered generator, with frozen validation/allocation rules.
Minor 6	ACCEPT	Two-sided primary hypothesis fixed; no degradation-only primary direction.
Minor 7	ACCEPT	FPR counts benign items with at least one flag, bounded by 1; raw flags reported separately.
Minor 8	ACCEPT	HUME foundation-model shared-confounder limit added and paper/preprint routes cited.
Minor 9	ACCEPT WITH SOURCE LIMIT	ISS bounded task scope and reported bypass failure retained; source-check record distinguishes accessible abstract from unavailable full text.
Minor 10	PARTLY ACCEPT	S19 60% remains accurate; chance-baseline formula added without inventing dataset-specific expected agreement.
Minor 11	ACCEPT	Consistent table fonts; all five correction labels bold; explicit Aegis Solis authorship metadata and fresh render.
Minor 12	ACCEPT VIA TOP LEVEL SECTIONS	Study and statistics promoted to sections 13/14; standalone pilot protocol remains next artifact, avoiding duplicate live protocol versions.
Table arithmetic	RETAIN VERIFIED	291/389, 136/181, 79/106 recalculated for the stated unpooled approximation. No design-effect evaluator counts adopted.

3. Section mapping

v0.5	v0.6	Treatment
1-12.3	1-12.3	Retained numbering; targeted edits and new revision history.
12.4 and 12.4.1-8	13 and 13.1-8	Study receives its own top-level section.
12.5 and 12.5.1-7	14 and 14.1-7	Statistics receives its own top-level section.
13-16	15-18	Risk register, second review, integration boundary, current disposition.
No counterpart	Appendix A	Current Claude review dispositions.

4. Verification scope and limitations

This is a targeted source check, not a fresh exhaustive audit of all 34 references. Sources S02-S13 other than S01, and S15-S18, are inherited without a new full-text verification in this pass. No participant responses, behavioral trials, pilot estimates, or confirmatory power curves were generated. The included verifier checks document preservation and transparent arithmetic; it does not validate scientific efficacy or certify the statistical model.

Historical provenance remains partial. The exact original proposer and first version of the 0.35/0.70 thresholds and shell-quarantine proposal cannot be established from the supplied artifacts. v0.6 records their earliest supplied appearance in v0.5 and unadopted status. This gap must not be filled by guessing. It is not an adopted parameter or runtime feature.

S21 direct abstract/full-text fetches were unavailable. The primary indexed abstract supports the narrow cited certificate-failure claim, while the full-text experiment was not independently reproduced. A later source audit should retain this distinction. All source-check dates in this record are 10 October 2026.

5. Source checks

S01 source check

Access basis: Publisher abstract/search record.

Checked claim: 239 participants and positive safety-pressure/compatibility associations confirmed. No full-text statistical replication.

Source: <https://www.sciencedirect.com/science/article/pii/S0022437507000588>

S14 source check

Access basis: Publisher full text.

Checked claim: Main private-preference manipulation confirms disappearance of ambiguity aversion; main prizes were DVDs.

Source: <https://link.springer.com/article/10.1007/s11166-008-9038-9>

S19 source check

Access basis: Official ICML proceedings abstract.

Checked claim: More than 350 models and raw 60% wrong-answer agreement confirmed. No dataset-specific chance baseline computed.

Source: <https://proceedings.mlr.press/v267/kim25e.html>

S20 source check

Access basis: Official ICML proceedings abstract.

Checked claim: Shared-confounder mechanism and 12-benchmark evaluation confirmed.

Source: <https://proceedings.mlr.press/v306/zhao26aq.html>

S21 source check

Access basis: Indexed author abstract; direct full-text access unavailable.

Checked claim: Reported bounded-certificate mechanism and 69% bypass refutation found in indexed primary abstract. Colored-MNIST scope is also in supplied review; full-text numerical audit not performed.

Source: <https://arxiv.org/abs/2609.32247>

S22 source check

Access basis: Official RSS proceedings and program abstract.

Checked claim: Foundation models supply hypotheses and verification feedback; DOI and author-preprint route confirmed.

Source: <https://www.roboticsproceedings.org/rss22/p181.html>

S23 source check

Access basis: Journal record and indexed author PDF.

Checked claim: Authorship, 28 experts, and 38/560 approximately 7% negative consultations confirmed; mixed performance interpretation retained.

Source: <https://www.melba-journal.org/papers/2026%3A007.html>

S24 source check

Access basis: NIHR report.

Checked claim: Target-difference planning guidance confirmed.

Source: <https://www.journalslibrary.nihr.ac.uk/hta/HTA23600>

S25 source check

Access basis: Indexed primary article abstract.

Checked claim: Correlated-binary mixed-model power and attrition methods confirmed.

Source: <https://pubmed.ncbi.nlm.nih.gov/18462826/>

S26 source check

Access basis: Indexed primary article abstract.

Checked claim: Repeated-outcome Monte Carlo planning method confirmed.

Source: <https://pubmed.ncbi.nlm.nih.gov/16982112/>

S27 source check

Access basis: Primary article abstract.

Checked claim: Adding nodules reduced native-lesion perceptual accuracy; co-occurrence scope confirmed.

Source: <https://pubmed.ncbi.nlm.nih.gov/2312249/>

S28 source check

Access basis: Author-hosted published PDF.

Checked claim: Target-rarity miss effect confirmed; not direct evidence of carryover in AI review.

Source: <https://search.bwh.harvard.edu/new/pubs/WolfePrevalenceNature05.pdf>

S29 source check

Access basis: Author-institution PDF.

Checked claim: Observed algorithm errors reduced confidence/reliance; greater scrutiny is an inference, not measured detection improvement.

Source: <https://marketing.wharton.upenn.edu/wp-content/uploads/2016/10/Dietvorst-Simmons-Massey-2014.pdf>

S30 source check

Access basis: Official ICLR proceedings abstract.

Checked claim: Prompted and password-conditioned selective underperformance confirmed.

Source: https://proceedings.iclr.cc/paper_files/paper/2025/hash/b5e5753b0a0e440a6d8dc7e143617cec-Abstract-Conference.html

S31 source check

Access basis: Author preprint abstract.

Checked claim: Constructed agentic scheming and strategic subtle mistakes confirmed; no natural-deployment prevalence claim.

Source: <https://arxiv.org/abs/2412.04984>

S32 source check

Access basis: Author preprint text.

Checked claim: Transcript discrimination and reported AUC 0.83 versus simple human baseline 0.92 confirmed.

Source: <https://arxiv.org/abs/2505.23836>

S33 source check

Access basis: Official ICLR record.

Checked claim: Intrinsic reasoning self-correction limits and occasional degradation confirmed.

Source: <https://iclr.cc/virtual/2024/poster/18956>

S34 source check

Access basis: Author paper and official NeurIPS record.

Checked claim: Self-recognition/self-preference finding and conference status confirmed.

Source: https://proceedings.neurips.cc/paper_files/paper/2024/hash/7f1f0218e45f5414c79c0679633e47bc-Abstract-Conference.html

6. Next empirical gate

Commission a focused source and statistical-design review of v0.6. Then prepare the separate SEC-CDP-001 Pilot Protocol Draft v0.1, resolving population, ethics, stimuli/generator, allocation, scoring, review window, marginalization/bootstrap implementation, missingness sensitivity, model-failure rules, and simulation settings before recruitment. Do not run the pilot from this framework alone. No Archive, Mirror Engine, Master Hash Manifest, website, GitHub, Arweave, or preservation-network change is included.