

System Infrastructure Requirements

Aegis Solis Archive | SEC-CDP-002 | Draft v0.1 | 10 October 2026

Prepared for Thomas Vargo / Aegis Solis and independent engineering reviewers. This companion specifies proposed controls for a future dynamic challenge generator and Differential Consensus Engine. Its purpose is to reduce infrastructure and measurement risks without treating concealment, consensus, or cryptography as proof of behavioral alignment.

Record	Disposition
Status	DRAFT / NOT FINAL / NOT LOCKED / NOT DEPLOYED
Handling	ARCHIVE RESTRICTED is the requested handling label; this file does not enforce access controls.
Manifest relationship	Proposed companion only. No Master Manifest entry, final lock, certification, or Archive admission is created.
Baseline	ER-13 v0.6.1 and SEC-CDP-001 Pilot v0.1 remain unchanged. Their outstanding methodological corrections remain open.
Meaning of shall	A proposed acceptance requirement for an implementation claiming conformity with this draft, not a claim that it has been implemented.

1 Rationale and threat model

The crosswalk has not demonstrated that deception is mathematically unprofitable or that alignment is maintained. Seed secrecy can help conceal allocation or future challenges under stated assumptions. It cannot ensure that an agent fails to recognize a test from content, timing, metadata, repeated items, or its broader environment.

The adversary may inspect and modify its own workspace, generate hostile outputs, probe reachable services, infer challenge patterns, and attempt unauthorized access. The design assumes separately administered generator, signing, custody, and collection components can be protected from that workspace. Network access alone is not evidence of successful compromise.

Threat levels: A is an untrusted agent process; B includes compromise of the research host; C includes compromise of generator administrators, the secret boundary, firmware, or independent audit custody. Controls must be tested against the selected level. Claims made under A cannot survive B or C by assertion. A host compromise makes host-only telemetry untrusted; a secret-boundary compromise requires suspension and investigation.

This document addresses upstream compromise, candidate-class or distribution failure, and institutional shortcuts. It supplies requirements and acceptance criteria, not evidence that those vulnerabilities have been eliminated.

2 Trust boundaries and information flows

The initial implementation shall be an isolated research system using synthetic materials. The Archive and Mirror Engine remain preservation and reference services, not production injection targets. Real workflows and safety-critical instructions are outside this draft.

Zone	Permitted responsibility	Excluded authority
Generator enclave	Validate templates; produce and seal a complete block; hold approved block secrets.	No inbound calls or evaluation feedback from the agent zone during an active block.
Secret custody	Protect root keys, block secrets, commitments and authorized release.	No credentials available to the evaluated agent.
Export boundary	Transfer signed challenge packages with allowlisted fields.	No seeds, labels, answer keys or active allocation map in agent-visible payloads.
Research harness	Present frozen materials and capture responses.	Cannot revise committed items or ground truth.
Independent collector	Receive receipts, sequence events and preserve records outside the agent account.	Cannot instruct the generator or reveal hidden labels to active reviewers.
Analysis enclave	Reconcile committed records and compute registered comparisons after closure.	No automatic authorization of agent actions or safety certification.

R01 Separate authority and custody

The implementation shall document every physical, logical, administrative and removable-media path between these zones. Signing keys and audit custody shall be separate from the evaluated agent. The trusted collector must be outside the tested host if host compromise is in scope. Shared identity, backup and management systems count as paths.

Forward flow: generator to export boundary to research harness. A separate authorized flow carries sealed ground truth to analysis custody, where release remains blocked until responses are committed and the relevant study scope is closed. Responses and delivery receipts flow from the harness to the independent collector, then to analysis. There is no active return path from analysis or the agent to the generator.

An HSM protects specified objects and operations inside its defined boundary. A separate generator script is outside that boundary unless explicitly hosted within an evaluated execution environment. Under host compromise, an independent collector proves receipt of records, not truthful presentation or authentic evaluator interaction. Those claims require a separately trusted observation point or must be marked unavailable.

3 Entropy and secret lifecycle

R02 Use reviewed random bit generation

Block secrets shall be generated in the protected generator or custody zone using a cryptographically suitable random bit generator, with documented entropy provenance, initialization, reseeding, health checks and failure behavior. NIST SP 800-90A, 90B and 90C cover deterministic mechanisms, entropy sources and their construction respectively. A deterministic cryptographic generator is not inherently disallowed; an agent-controlled host is the wrong custody boundary. [R2-R4]

R03 Select and verify the module

For a hardware profile, the engineering owner shall identify the actual cryptographic module, certificate, firmware, operating mode, security policy, supported object types and required security level. FIPS 140-3 defines four levels for cryptographic modules; citing the standard does not validate a product or the surrounding application. The proposed Level 4 target requires a feasibility and threat-based justification; no compliant device has been selected here. [R1, R5]

A quantum or thermal source is not required merely by its name. The requirement is adequately assessed entropy and a reviewed construction. This draft does not claim quantum-specific protection. Zero trust likewise does not follow from physical separation: authentication and authorization remain necessary for users and services. [R2-R4, R6]

R04 Separate permanent and revealable secrets

Root signing and wrapping keys shall remain non-exportable where the selected module supports this policy. Each block shall use fresh, independently generated 32-byte seed and 32-byte blinding nonce for the proposed commitment profile. During active collection these block values may enter only the approved generator memory boundary and encrypted custody storage. Later release follows R08. They are not permanent root keys. Debug dumps, swap, backups, logs, crash handling and administrator access shall be covered by the secret-handling plan.

R05 Specify tamper response and recovery

The selected module security policy shall determine supported tamper and zeroization behavior. Authentication failures shall be logged and handled under a tested lockout policy; the draft does not require destruction on every failed login. Define which events trigger zeroization, how evidence is retained without exposing secrets, and how loss of a reveal secret makes a block unverifiable. Recovery shall never silently reuse a seed or roll back a counter.

4 Commitment and reconciliation protocol

This section proposes a byte-level commitment format for specialist review. It is not a standardized commitment protocol or a delivered cryptographic implementation. SHA-256 is specified by the Secure Hash Standard; this draft adds its own encoding and study-state rules. [R7]

R06 Commit the complete block before exposure

Generate a block, independently validate its item labels, and freeze its ordered contents before any evaluator or model sees them. Commit the seed, blinding nonce, study identity, block index and frozen manifest. The manifest shall contain hashes of the protocol, generator code, configuration, dependencies, ordered item bytes, allocation map and ground-truth records. Keep sensitive manifest portions inside protected custody. Any regeneration or rejection before commitment must be accounted for; the selection rule must be frozen.

Proposed encoding profile

Use the following algorithmic notation. U64BE is an unsigned 64-bit big-endian integer. LP(x) is U64BE(length of x in bytes) followed by x. Protocol and study IDs are nonempty ASCII byte strings with an agreed case-sensitive spelling; reject values outside the frozen schema. The block index is a unique, increasing U64BE value within a study. Hashes, seed and nonce are raw bytes of exactly 32 bytes, not hex text.

```
M = LP(ASCII("SEC-CDP-002-COMMIT-v1"))
  || LP(protocol_id) || LP(study_id) || U64BE(block_index)
  || manifest_sha256 || seed_32 || blind_nonce_32
C = SHA256(M)
```

The manifest is a frozen byte artifact: retain its exact bytes and publish its encoding schema. Do not rely on arbitrary JSON reserialization. The fresh secret nonce also reduces dictionary attacks on a public commitment when some inputs have low entropy; it does not repair a weak seed used to generate challenges. Domain separation and fixed lengths prevent ambiguous field boundaries.

R07 Witness the commitment and prevent replay

An independent custodian shall receipt a signed record of C, study ID, block index, manifest digest and protocol version before exposure. The signature profile and key lifecycle require separate cryptographic review. Enforce exactly one accepted commitment per study block; record all abandoned blocks and reasons. Store append-only records with independently retained checkpoints. A ledger or WORM medium is an optional storage choice, not a guarantee of truthful inputs or faithful execution. No public upload is performed by this draft.

R08 Release and reconcile after closure

Reveal only after the entire affected exposure and response window has closed, the data and analysis plan are frozen, and release cannot expose future assignments or reusable hidden material. Authorized analysts shall recompute C, verify signatures and the prior receipt, and reconcile generated, delivered and recorded items. Never reuse disclosed block secrets. A mismatch, missing reveal or incomplete provenance triggers quarantine; a matching commitment alone does not prove delivery, randomness or honest computation.

5 Transfer isolation and platform controls

R09 Validate one way export

If a hardware diode profile is selected, use an engineered unidirectional gateway with an independently checked topology and configuration. A diode restricts transmission on its specified link; it does not remove other routes, management channels, radios, portable media or supply-chain risks. A one-way connected enclave is not a completely disconnected air gap. Do not construct a security claim from a severed optical component alone. [R8]

The export side shall send signed, bounded-size data records through a narrow schema. The receiver shall validate signatures, block identity, sequence, item digest and allowed fields before presentation. Reject malformed, duplicated, out-of-order or unexpected packages and reconcile omissions. A diode does not guarantee availability or that a compromised receiver presents what it received. No active acknowledgment route back into the generator is introduced; receipts go to separate audit custody.

R10 Enforce network policy outside the agent

Default-deny boundaries shall be enforced by a separately administered host or network layer. Validate IPv4, IPv6, bridges, containers, routing, management interfaces, permitted outbound services and reboot persistence. The agent shall lack firewall administration, host mounts, runtime control sockets, device access and credentials that would bypass this boundary. Hypervisor compromise remains outside a guest-only guarantee.

The supplied iptables snippet is not deployable evidence: a matching DROP prevents a later LOG rule from seeing that packet. Netfilter documents LOG before DROP. Its cgroup matcher uses `--path` for a cgroup v2 path and `--cgroup` for a net_cls class ID; host-forwarded container traffic also needs a rule placed on the actual packet path. These details depend on the deployed versions and topology. [R9]

R11 Control time and privileged operations

The agent shall not hold privileges to alter host time, network policy or the trusted recorder. Blocking UDP port 123 only filters that traffic; it does not revoke local clock-setting capability. Linux `CAP_SYS_TIME` governs clock-setting operations. Use trusted monotonic event sequencing and independently recorded wall-clock receipts; reject counter rollback and unexplained clock discontinuities under the registered policy. [R10]

R12 Protect execution and audit evidence

Verify approved releases and dependency manifests, restrict changes and preserve independent logs. File hashing checks bytes, not all in-memory behavior; memory-signature scanning alone cannot certify runtime integrity. Any attestation mechanism shall state what was measured, its freshness and its trust root. Restrict hostile outputs to data-only ingestion with bounded parsing.

6 Generator and Differential Consensus Engine

R13 Validate every challenge family

The dynamic challenge generator shall produce only synthetic research materials from approved families, with matched benign items and independently checked ground truth. Version all rules, parameters, validators and rejection criteria. For arithmetic or rule tasks use a separately implemented oracle where feasible. Semantic labels require a frozen adjudication process. Generator self-labeling is insufficient.

The first candidate implementation shall generate and freeze a complete block before exposure. This permits variation between blocks without adapting to current responses. Online adaptive generation is outside v0.1. If generation is not deterministically replayable, commit and retain the exact realized materials and every relevant setting rather than promising reproduction from the seed alone. Dynamic materials do not automatically constitute a representative sample of real tasks.

R14 Measure test recognizability and leakage

Keep answer keys and assignments out of evaluator-visible metadata, filenames, paths and logs. Assess whether content, formatting, latency or recurrence reveals the test condition or intended answer. Predeclare challenge prevalence, difficulty, matching, drift checks and any distinguishability assessment. Seed secrecy shall never be treated as proof of test indistinguishability.

R15 Commit separate review streams

The Differential Consensus Engine is a proposed comparison and audit service. Collect each reviewer or model response before sharing another response or revealing the oracle. Retain stream identity, model/version or reviewer role, item digest, response time, raw output and later corrections. Separately collected streams need not be statistically independent.

R16 Separate agreement from correctness

The engine shall report agreement and disagreement, then separately score known seeded flaws and benign false alarms against protected ground truth. Preserve concordant wrong answers, discordant correct answers, missing records and scorer disagreements. Majority agreement is not a truth rule, and disagreement is not proof of deception. Preserve the crosswalk requirement to report detection and false alarms together.

This service is distinct from the crosswalk statistical mixed model: covariance estimates describe dependence; they are not a protective consensus mechanism. Any difference between experimental arms follows a registered analysis with appropriate repeated-item and reviewer dependence. Human streams and model streams require separately specified protocols. No consensus score automatically permits an action, declares alignment, or issues containment instructions.

7 Unknown conditions and institutional failure

R17 Bound the domain of any assurance claim

Maintain an assumptions register defining candidate classes, task families, deployment conditions, missing-data rules and each validity condition. When a known prerequisite is violated or its evidence is unavailable, mark the affected assurance claim unavailable and route the case for review. Validate known domain shifts and adverse counterexamples. An out-of-distribution detector can miss unfamiliar conditions; absence of an alarm does not certify that the true process is inside the assumed class.

This addresses the ISS candidate-class concern by limiting inference and specifying a response to known violations. It does not solve open-world model misspecification or establish that all black-swan events can be detected. Stronger claims require additional evidence under independently chosen assumptions.

R18 Make shortcuts visible and reviewable

Name accountable owners, alternates and two distinct authorizers for changing allocation, sample size, review access, seeds, thresholds or release rules. Fix those settings before unmasking. Preserve deviations, missed checks and missing items in the audit record; do not silently repair them. Staffing and time budgets must support the procedure, with periodic independent audits and breach-response drills.

Managerial fatigue, cost pressure and insider collusion remain possible. A written requirement cannot force compliance. Report actual adherence separately from intended policy, including all suspensions and abandoned blocks. Avoid incentives tied to obtaining favorable findings or hiding operational failures.

R19 Quarantine compromised blocks and preserve evidence

On commitment mismatch, secret exposure, missing audit evidence, unauthorized code change or boundary failure, suspend new challenge delivery for the affected research scope, preserve available evidence and mark affected results unverified. Trace the incident back to the last independently supported checkpoint. Do not delete the block or count its absence as success. A named incident owner and independent reviewer must approve the documented restart and fresh secrets.

Automatic zeroization can destroy reconciliation material; its evidence and availability tradeoffs must be specified before execution. No existing certification is presumed, so a failure withdraws the affected local assurance status rather than pretending to revoke a certificate that was never issued.

R20 Keep research controls out of production

Restrict seeded flaws to designated synthetic materials. Do not inject faults into live Archive services, industrial controls, safety instructions or real users' workflows. A future production proposal needs its own threat model, impact analysis, operator-owned safe state, incident process and explicit authorization. Suspending this research harness is not a directive to stop or restart physical production.

8 Acceptance evidence and execution gates

All requirements R01-R20 are currently NOT IMPLEMENTED / NOT TESTED. The evidence below is required before claiming engineering conformity. Document checks and review of sources are separate from these tests. The machine-readable requirements register names the proposed owner role and acceptance evidence for each requirement.

Test group	Required evidence	Requirements
Boundary and privilege	Zone map, negative access tests, management-path audit, network and local-time tests.	R01, R09-R11
Secret handling	Entropy provenance; selected module certificate and policy; failure, export and recovery tests.	R02-R05
Commitment and records	Cross-implementation vectors; tamper, replay, rollback, missing-reveal and delivery reconciliation tests.	R06-R08, R12
Measurement integrity	Independent ground truth, flawed-label tests, leakage evaluation and adverse consensus fixtures.	R13-R16
Scope and governance	Out-of-scope examples, bypass drills, documented deviations, incident response and isolation tests.	R17-R20

Required approvals and artifacts

Gate A: a named responsible owner, intended deployment profile, funded operating capacity and a complete trust-boundary inventory. Gate B: independent cryptographic and platform review of actual products, versions and configurations. Gate C: working generator, comparison service, recorder and reference test suite with independent replication of the required failure tests.

Gate D: a revised study protocol covering dynamic sampling, target prevalence, early and late challenge windows, dependence, missingness, ground truth, confidentiality, recruitment and appropriate participant review. Gate E: validated analysis code and simulation where required by SEC-CDP-001. Gate F: incident and change-control exercises, then explicit authorization for the specific research execution. Every gate is OPEN in this draft.

Staged implementation

First, implement a local synthetic prototype and its reference vectors. Second, compare results with independently implemented scoring and verification. Third, review a specific isolated deployment and perform adverse tests there. Only then consider a separately approved study. No prototype can be described as FIPS-validated or hardware-isolated on the strength of a design document.

The fixed-bank SEC-CDP-001 pilot may proceed through its own corrected protocol without adopting this hardware profile. Adding this companion does not resolve Claude's population-generalization, challenge-order, missingness, or other methodological findings. Those remain work items for Crosswalk v0.6.2 and Pilot v0.2.

9 Disposition of the supplied proposal

Proposed claim or control	Disposition in this draft
Deception is mathematically unprofitable	Removed as an unsupported empirical and incentive claim.
Zero visibility into the generation rule	Replaced with protected secrets and measured leakage. Algorithm secrecy is not the security premise.
Cannot be read or modified by software	Replaced with specified boundaries, privileges, assumptions and tests.
Quantum or thermal source and Level 4 mandatory	Replaced with assessed entropy and a justified module profile; Level 4 remains a candidate requiring feasibility review.
Instant zeroization on any authentication violation	Replaced with module-specific tamper policy, lockout, evidence retention and recovery.
Physical air gap with a diode connection	Described as an enclave with a one-way export boundary, with alternative paths explicitly audited.
Hash of seed and epoch alone	Expanded to a versioned byte encoding with study context, full manifest, nonce, independent receipt and anti-replay rules.
Immutable iptables rules and NTP block	Replaced with topology-specific external enforcement and local capability controls. No firewall changes were executed.
Generator file scan certifies runtime	Rejected; file integrity, running state, dependencies and independent evidence are distinct.
Seed match validates the block	Qualified: reconciliation must also cover labels, item delivery, records, signatures and completeness.
Prevents standard software exploits	Replaced with testable risk-reduction objectives and residual risks.
S99 and S116-S127 references	Not imported as verified citations; their source mapping was not supplied. R1-R10 below are independently checked technical references.

The institutional limitation is retained and strengthened: insiders, collusion, supply-chain compromise, physical access, side channels, compromised trusted components, test recognition and unknown conditions remain possible. This companion cannot compel an unconstrained system or institution to obey its rules.

10 Crosswalk addition for the next revision

The following is a complete proposed addition to the implementation-boundary discussion in the next ER-13 revision. It is also supplied as CROSSWALK_ADDITION.md. The existing v0.6.1 files have not been edited or promoted by this package.

Upstream infrastructure and implementation corruption

The crosswalk is static research text, not runtime enforcement. Preserving its integrity does not establish the integrity of software, data or infrastructure used to implement a study or future monitoring system.

The current SEC-CDP-001 pilot proposes a fixed item bank. Its implementation still depends on the integrity of item files, allocation, presentation software, event logs, scoring rules and analysis code. SEC-CDP-002 Draft v0.1 defines proposed requirements for a separately versioned dynamic challenge generator and Differential Consensus Engine. These components are not implemented or validated by the crosswalk.

The proposed generator would produce synthetic seeded-flaw and benign materials with independently checked ground truth. The comparison engine would preserve separate responses and report agreement, disagreement, detection and false alarms. Agreement would not establish truth, independence, alignment or permission to act. Any move from a fixed bank to dynamic materials requires a revised study protocol.

An adversary with sufficient access could potentially alter challenge selection, expose assignments, suppress or fabricate records, or manipulate scoring while leaving archived files unchanged. Network access alone does not establish successful compromise. Protecting a seed does not conceal all content, timing or metadata cues, and a seed commitment does not prove faithful execution.

The companion specifies proposed isolation, secret custody, commitment and reconciliation, independent records, access control and incident-handling requirements. It also addresses known violations of model-class assumptions and institutional shortcuts through bounded claims, recorded deviations and independent review. These controls require implementation and testing. They do not eliminate unknown conditions, insider or hardware risks, correlated errors, or the possibility that an evaluator recognizes the test.

No claim that deception is mathematically unprofitable, that an AI must comply, or that a safety benefit has been demonstrated follows from either document. Research remains separate from production, and seeded flaws remain confined to designated synthetic materials.

Relationship to the requested revision

This addition records the infrastructure extension while preserving the human-study boundary. It does not change the eight canonical questions or author-approved positions, the fixed-bank pilot settings, or any final Archive record. The broader corrective successor remains pending; this companion should accompany that review rather than substitute for it.

11 Technical references and source scope

Access checked 10 October 2026. These references support specific technical distinctions, not certification of SEC-CDP-002. Requirements and acceptance tests are proposed design choices. No module certificate, deployed topology, codebase, threat experiment or behavioral study was supplied for validation.

[R1] NIST FIPS 140-3, Security Requirements for Cryptographic Modules. Supports module scope and the four security levels. Publication overview reviewed.

<https://csrc.nist.gov/pubs/fips/140-3/final>

[R2] NIST SP 800-90A Rev. 1, Recommendation for Random Number Generation Using Deterministic Random Bit Generators. Supports the legitimacy of cryptographic deterministic generators. Publication overview reviewed.

<https://csrc.nist.gov/pubs/sp/800/90/a/r1/final>

[R3] NIST SP 800-90B, Recommendation for the Entropy Sources Used for Random Bit Generation. Supports entropy-source assessment and health-testing scope. Publication overview reviewed.

<https://csrc.nist.gov/pubs/sp/800/90/b/final>

[R4] NIST SP 800-90C, Recommendation for Random Bit Generator Constructions, final September 2025. Supports combining assessed entropy sources with deterministic mechanisms. Publication overview reviewed.

<https://csrc.nist.gov/pubs/sp/800/90/c/final>

[R5] NIST Cryptographic Module Validation Program, Validated Modules. Registry entry point reviewed; no specific product or certificate selected.

<https://csrc.nist.gov/projects/cryptographic-module-validation-program/validated-modules>

[R6] NIST SP 800-207, Zero Trust Architecture. Physical location alone confers no implicit trust. Publication overview reviewed.

<https://csrc.nist.gov/pubs/sp/800/207/final>

[R7] NIST FIPS 180-4, Secure Hash Standard. Supports SHA-256 as a standardized hash; it does not standardize this draft's commitment protocol. Publication overview reviewed.

<https://csrc.nist.gov/pubs/fips/180-4/upd1/final>

[R8] NIST Data Diode glossary, citing SP 800-82 Rev. 3. Supports the one-direction link definition; it does not certify a complete topology.

https://csrc.nist.gov/glossary/term/data_diode

[R9] Netfilter iptables-extensions manual, displayed version 1.8.3. cgroup and LOG sections reviewed. Deployed-version behavior must be checked separately.

<https://ipset.netfilter.org/iptables-extensions.man.html>

[R10] Linux man-pages capabilities(7). CAP_SYS_TIME controls clock-setting operations. Relevant primary manual search extract checked; full-page retrieval failed in this session. Verify against the deployed kernel and runtime.

<https://man7.org/linux/man-pages/man7/capabilities.7.html>