System Infrastructure Requirements

Aegis Solis Archive | SEC-CDP-002 | Draft v0.1.3 | 11 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	Coordinated with ER-13 v0.6.4 and SEC-CDP-001 Pilot v0.4. Prior v0.6.1 and v0.1 packages are preserved unchanged.
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.
Evaluated agent zone	Receives one presentation unit at a time through a data-only interface.	No access to harness storage, later items, other streams' responses or timing, collector interfaces, labels or allocation.
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. The harness authenticates each stream, releases only its current authorized unit and accepts a bounded response; it never mounts its block store or audit credentials into the agent zone. Prior items may remain in that stream's context only under its frozen context policy. Stream isolation and read-ahead denial require negative tests, including host-compromise limits. The harness-side block store contains unlabelled presentation units only. Labels, answer keys and labelled blocks follow a separate protected custody route and never enter that store. Validation and post-response adjudication use separate authorized custody, not the presentation harness.

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 adequate 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. The plaintext asset inventory in section 3A is mandatory. In replayable generation, a seed with known generator/configuration and its labelled block are equivalent for answer secrecy. The HSM security level is justified by the keys and operations it actually protects; it does not protect exported seeds or labelled blocks in general host memory.

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.

3A Plaintext asset inventory

This inventory is an acceptance requirement, not an assertion of implemented isolation. Plaintext processing windows shall have named owners and concrete retention limits before execution. Storage, backups and transfer outside these windows are encrypted under separate custody; prohibit secrets in logs, dumps, swap and shared caches, or protect those paths explicitly.

Asset	Permitted plaintext location	Lifetime and release
Root signing and wrapping keys	Inside the selected module boundary only.	Key lifecycle; no plaintext export. Rotation and module recovery policy apply.
Block seed	Protected generator memory and authorized reveal/analysis memory.	Generation or approved replay window; otherwise encrypted custody. Reveal only under R08; then never reuse.
Block nonce and component salts	Protected generation/commitment and reveal memory.	Secret through R08; encrypted at rest, retained for audit opening. Never shared with the agent.
Unpresented items (unlabelled)	Generator, protected custody and harness-side block store; never mounted into the agent zone.	Protected until authorized per-unit presentation. Keep each unit and its signature confidential until that unit is released.
Labels, answer keys and labelled block	Generator, authorized validator/adjudicator and analysis custody only; never exported to the harness side.	No access by active streams or presentation harness. Authorized release after all affected windows close under R08; governance controls remain.
Allocation map	Allocation custodian and minimal harness routing memory.	Never visible to evaluated streams or blinded scorers; unblind under frozen protocol after closure.
Manifest and manifest digest	Protected commitment custody and authorized analysis memory.	Secret until R08. No raw or component digest in public receipts or agent-visible metadata.
Uncommitted responses	Originating stream and access-controlled capture memory.	No cross-stream visibility. Commit before comparison; collector may retain sealed records under separate custody.

Cryptographic release approval does not authorize publication of identifiable participant data. Post-study retention and authorized access remain subject to the study governance plan. The agent knows its own current response and any permitted prior context; the boundary prevents access to other streams, not self-knowledge.

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. Before reveal, only the blinded commitment C and approved non-sensitive receipt metadata leave protected commitment custody. The manifest digest remains secret until R08. This restriction concerns commitment records; authorized challenge presentation still follows the separately controlled export path.

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-v2"))
  || 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 receipt custodian outside the agent account shall receipt a signed record containing C, non-sensitive study ID, block index, protocol/profile version, signing-key ID and receipt time before exposure. Do not include the manifest, manifest digest, component digests, salts, seeds, labels or allocation in this record. The public receipt service is distinct from protected secret custody even if an organization operates both. Enforce one accepted commitment per study block, preserve abandoned blocks and reasons, and retain independent checkpoints. All metadata and identifiers shall be reviewed for condition leakage. WORM storage or a ledger does not establish truthful inputs or faithful execution.

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. Authorized opening includes the exact manifest bytes and every required component salt, with recomputation of all digests before C. Release may be restricted to authorized analysts; do not publish participant identifiers or protected responses merely to open a commitment. At least one verifier independent of the study team shall open the commitment, recompute every required component digest and C, check the original receipt and record the result with the receipt custodian. Authorized access alone does not establish independence. Appoint the verifier and document conflicts before exposure; require necessary protected access and log release without publishing participant data. Without this independent opening record, mark reconciliation incomplete and keep the block unverified.

4A Secret component hashing and leakage tests

For each manifest component containing low-entropy or secret material, including labels, answer keys, allocation and unpresented item content, use a fresh independent 32-byte secret blinding salt. Salts are not ordinary public password salts. They remain secret until the authorized R08 opening; do not reuse them or derive them from a revealed seed.

```
D_x = SHA256(LP(ASCII("SEC-CDP-002-COMPONENT-v1"))
    || LP(study_id) || U64BE(block_index)
    || LP(component_id) || salt_x_32 || LP(component_bytes))
```

LP and U64BE follow section 4. Component IDs are unique within a block; retain exact byte encodings. The protected manifest contains D_x in place of a bare digest for each secret component. Public non-secret code/version files may retain ordinary integrity hashes. The full manifest digest is still withheld, even when its secret components are blinded. A separate salted opening record is protected and retained for R08 reconciliation.

R07 acceptance evidence shall include attempts to infer labels and allocation from the complete receipted record and all agent-visible metadata, given the public generator/configuration and visible items. Test small candidate spaces, component-digest leakage, public or reused salts, accidental manifest export, metadata cues, receipt replay and premature release. Use intentionally vulnerable synthetic fixtures as negative controls. These engineering tests remain NOT RUN; a source-code or document check does not prove hiding.

Risk qualification: a bare digest can become an offline guessing oracle when an adversary can reconstruct candidate manifest bytes. Knowledge of the algorithm alone does not establish that all other manifest inputs are known. The reviewer's candidate counts are combinatorial illustrations, not measured recovery costs or demonstrated compromise. Secret randomness and custody address this channel under cryptographic and implementation assumptions; they do not conceal flaws evident in the challenge content itself. [R11]

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, unit signature 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. The per-unit integrity check verifies a signature over the exact unit bytes and their unambiguous study/block/unit context; no bare content digest of an unpresented item is exported. Signatures and signed units stay in the protected harness store until per-unit release, never in public receipts or agent-visible previews. A signature authenticates bytes; it is not a hiding mechanism. Test rejection of changed units, context substitution, replay and premature metadata disclosure.

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. Assign a protected pair/template identifier. Never show both a seeded-flaw item and its corrected counterpart, or a revealing duplicate, to the same model stream or human participant within a study. Enforce this across blocks and retained context; a declared fresh/reset session counts separately only if isolation is established. Audit assignment history without exposing pair IDs to evaluated streams.

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. Test paired-variant leakage and read-ahead access as well as metadata cues.

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. For each model stream retain exact prompts/instructions, model/version, decoding settings, provider sampling seed when available (otherwise explicitly unavailable), tool access/results and the context/reset policy. This sampling seed is distinct from the secret challenge seed. Shared context makes carryover part of the stream; preserve sequence and do not treat items as independent fresh calls. Keep other streams' responses and timing concealed until all corresponding responses are committed; a stream cannot access collector or harness storage.

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. If people or their data are affected, notify the responsible investigator, ethics body and data-protection contact where applicable, under the approved incident plan. End affected sessions with appropriate debriefing and agreed payment; preserve privacy and record any justified delay in disclosure.

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. Agent read-ahead, stream-isolation and paired-variant tests; asset lifetime/custody audit. Verify unlabelled-only harness export and confidential unit/signature staging.	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. Receipt-based label/allocation guessing and secret component-salt leakage tests. Independent verifier appointment, opening record and receipt-custodian acknowledgment.	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.

SEC-CDP-001 Pilot v0.4 retains the fixed bank and does not require this hardware profile. Crosswalk v0.6.4 and Pilot v0.4 address the latest review findings. Both require independent review and completion of their own execution gates. This infrastructure companion does not establish scientific validity.

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 Integrated crosswalk addition

The following text is integrated into section 12.4 of Crosswalk v0.6.4 in this coordinated package. CROSSWALK_ADDITION.md retains the same text for review. Original v0.6.1 files are preserved unchanged.

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 v0.1.3 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 coordinated successor preserves the eight canonical questions and author-approved positions. The human pilot remains fixed-bank. The revised secrecy and agent-boundary requirements remain unimplemented and untested.

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. These inherited access notes describe the earlier source checks; the v1.2 correction record states this round's limited source scope.

[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

[R11] Daniel Gottesman, CMSC/Math 456 Cryptography, Lecture 19 (Fall 2022), slides 16-18. University of Maryland. Hash commitments with random opening material and the distinction between hiding and binding. Lecture text checked 10 October 2026; this is background, not a security proof or endorsement of this proposed profile.

https://www.cs.umd.edu/class/fall2022/cmsc456/lecture19.pdf