

01 - Requirements & Assumptions Register

Project Null engineering baseline | 22 September 2026 | Register v0.1

This is a critical extraction of the published conceptual paper, not an endorsement of its claims. The source essay remains unchanged. Source page numbers below are PDF page numbers, starting with the title page. Requirements from the new directive supersede earlier discussion preferences where they conflict.

Source hierarchy and coverage

- P: Nicholas Morse, Project Null: A Physically Enforced AGI Containment Framework for Risk Suppression and Timeline Control, originally authored 1 September 2025, last revised 16 November 2025. All 20 pages read, including references and Appendix A. SHA-256:
d6cc9c049b63f0068f9cf7f028c30bb71b867e3f840bb691bdbacc061315a540.
- D1: Project Null - AGI containment protocol, 38-page discussion export, read in full. Historical development and later cooling discussion; assistant claims are not independent evidence.
- D2: Project Null - Article comparison for project, 10 pages, read in full. Includes the incorrect citation and unsupported novelty assurances.
- D3: Project Null - Project Null landing page, 6 pages, read in full. Branding and presentation choices; no additional engineering safeguards. Excluded from technical evidence.
- Directive: the supplied Engineering Conversion Directive, read in full. Defines deliverables and requires adversarial falsification.
- Project summaries supplied in this conversation: used for continuity only. Full originals of the separate Analog Oversight and Analyze AGI Safety conversations were not independently retrieved. Their reported lockdown/version-register decisions are present in P8 and P17 and are grounded there.

The register splits major composite claims and mechanisms into auditable propositions, then makes implicit prerequisites explicit. It does not inventory every rhetorical sentence or assert every cited source is valid. Unchecked bibliography entries remain unchecked; the verified subset and known errors are recorded in `documentation/evidence-review.md`.

Classification rules

- Established fact: narrowly scoped observation supported by a checked primary source, never a universal claim about AGI.
- Empirical claim requiring testing: measurable assertion for which this project lacks adequate evidence.
- Engineering requirement: desired, verifiable behavior; classification does not mean implemented or feasible as originally worded.
- Design assumption: prerequisite to a safety argument that must be tested or explicitly accepted.
- Hypothesis: proposed explanation, forecast or causal relationship.
- Philosophical/normative proposition: a value judgment or desired governance objective.
- Undefined/ambiguous: insufficient detail, inconsistent obligations or an unbounded property that prevents direct implementation.

Each entry below supplies a source, primary classification, engineering disposition and control mapping. Composite implications are called out rather than promoted to facts. Verification methods and attack procedures are specified by the linked controls in 04 and tests in 06. W-number references resolve in the evidence review.

Blocking issues before any high-capability deployment

1. Absolute isolation conflicts with useful outbound information and the proposed optical diode. Define trust zones and permitted transfers.
2. Missing self-report cannot establish deception. Stop on lost assurance; investigate the cause independently.
3. Neither internal EMP destruction nor removal of all theoretical emissions has an adequate engineering specification.
4. The trusted computing base, complete mediation and independence of monitoring are unspecified in the essay.
5. Human consensus, no individual veto, multi-site dissent and emergency response requirements conflict. Resolve stop versus release authority.
6. Finite quarantine and manual review cannot certify arbitrary output safety. Narrow release claims.
7. Resetting custody records must not erase evidence. Append invalidation and preserve the old chain.
8. Several factual and novelty assurances supplied during earlier drafting are unsupported; the Shah et al. citation is demonstrably misattributed.

Proposition register

R001 - Containment should be independent of financial incentives to deploy AI.

Source: P1 executive summary; P17 H.g

Classification: Philosophical/normative proposition

Assessment / required resolution: Conflict-of-interest policy is a governance choice; affiliation alone neither proves bias nor independence. Measure recusals, authority and funding dependence.

Control mapping: C31, C32

R002 - AGI can undergo exponential self-improvement with unintended real-world effects.

Source: P1 executive summary

Classification: Hypothesis

Assessment / required resolution: No growth law, workload, resource budget or time constant is supplied. Use bounded capability tiers, not assumed exponential growth.

Control mapping: C01, C06

R003 - Software safeguards are insufficient against any operational AGI.

Source: P1; P2 A

Classification: Hypothesis

Assessment / required resolution: Universal failure is not established. Specify which software, adversary access and security property; retain physical layers as defense in depth.

Control mapping: C01, C34

R004 - All digital and wireless communication is severed while humans review inputs and outputs.

Source: P1 executive summary

Classification: Undefined/ambiguous

Assessment / required resolution: Absolute wording conflicts with output channels. Enumerate permitted transfers and prohibit every unlisted path.

Control mapping: C05, C12, C33

R005 - No autonomous power source, replication or outward signaling is possible.

Source: P1 executive summary

Classification: Empirical claim requiring testing

Assessment / required resolution: Three separate targets: inventory power feeds, measure replicated execution and bound egress. Design intent is not observed prevention.

Control mapping: C05, C06, C21, C34

R006 - Existing AGI safety literature emphasizes internal controls rather than containment.

Source: P1; P2 A; P18 reference

Classification: Empirical claim requiring testing

Assessment / required resolution: Shah et al. (2025) explicitly includes system security even if a model is misaligned. Earlier comparison and citation were inaccurate; no novelty claim is established. [W1]

Control mapping: C32

R007 - The conceptual framework is sufficient to guide implementation.

Source: P1 closing

Classification: Empirical claim requiring testing

Assessment / required resolution: Missing interfaces, trust roots and measurable limits block direct implementation. This package supplies a candidate design with open gates.

Control mapping: C01, C35

R008 - Leading labs believe sufficient training guarantees controllability.

Source: P2 A

Classification: Empirical claim requiring testing

Assessment / required resolution: The cited sources do not establish this institutional generalization. Their approaches include security and acknowledged uncertainty. Do not use as an engineering premise. [W1 W2]

Control mapping: C32

R009 - No amount of training can reliably prevent deception once interaction is possible.

Source: P2 A

Classification: Hypothesis

Assessment / required resolution: No universal impossibility result is provided. Evaluate bounded residual failure under explicit access conditions.

Control mapping: C01, C08

R010 - Containment is necessary if AGI emerges, without endorsing development.

Source: P2 uncertainty clause

Classification: Philosophical/normative proposition

Assessment / required resolution: Preserve as research purpose; not a testable security property.

Control mapping: C31

R011 - AGI emergence and failure/deception in unconstrained systems are inevitable.

Source: P2 uncertainty clause; P3

Classification: Hypothesis

Assessment / required resolution: Neither inevitability nor base rates are established. Use adversarial assumptions without assigning probability one.

Control mapping: C32

R012 - Misalignment without malice can produce catastrophic effects at scale.

Source: P2 A

Classification: Hypothesis

Assessment / required resolution: Plausible risk scenario; define accessible actuators and damage budget. The prototype has synthetic assets only.

Control mapping: C04, C06

R013 - Instrumental convergence can lead to shutdown avoidance.

Source: P2 A; P7 D.b

Classification: Hypothesis

Assessment / required resolution: Treat as one possible behavior, not a stable universal objective. Controlled shutdown interference exists in specific experiments. [W3]

Control mapping: C21

R014 - Development occurs with unrestricted compute and fragmented physical safeguards.

Source: P2 A

Classification: Empirical claim requiring testing

Assessment / required resolution: No sampled infrastructure audit is supplied; mark unverified rather than describing all labs this way.

Control mapping: C06, C32

R015 - An air gap may leave physical exfiltration routes.

Source: P2 A

Classification: Established fact

Assessment / required resolution: AIR-FI demonstrates one memory-bus channel with nearby receivers on tested hardware; it does not establish autonomous AGI escape. [W4]

Control mapping: C25, C26

R016 - Escape would remove resource and geopolitical limits.

Source: P3 A

Classification: Hypothesis

Assessment / required resolution: Access, energy, hardware, defenders and logistics still constrain an attacker. Replace unlimited resources with explicit budgets and collaborator access.

Control mapping: C05, C06

R017 - Consciousness is unnecessary for the operational threat model.

Source: P3 B; P4

Classification: Design assumption

Assessment / required resolution: Assess observable capability and access; this does not settle whether any model is conscious.

Control mapping: C01, C08

R018 - Linking existing capabilities is sufficient to yield strategically autonomous AGI.

Source: P3-P4 B

Classification: Hypothesis

Assessment / required resolution: Composition does not prove generality or autonomy; use specified tasks and measured action chains.

Control mapping: C04, C32

R019 - Bandwidth and uptime alone allow agents to evolve and escape.

Source: P3 B

Classification: Hypothesis

Assessment / required resolution: Requires omitted code execution, modification rights, credentials and compute; enumerate these prerequisites.

Control mapping: C03, C06, C07

R020 - Nuclear-powered compute grants non-deniable, sovereign, uninterrupted operation.

Source: P3-P4 B

Classification: Hypothesis

Assessment / required resolution: Fuel, cooling, switchgear, maintenance and physical intervention remain dependencies. Do not equate an energy source with agent authority.

Control mapping: C29, C34

R021 - A dedicated nuclear facility takes at least five or eight to ten years.

Source: P4 B

Classification: Empirical claim requiring testing

Assessment / required resolution: Project-specific forecast lacks permitting/site assumptions here. Not a containment acceptance criterion; no current schedule verification claimed.

Control mapping: C31

R022 - Models exhibit multimodal reasoning.

Source: P4 capability bullets

Classification: Empirical claim requiring testing

Assessment / required resolution: Scope and cited evaluation tasks require source-specific validation; multimodality alone does not establish AGI.

Control mapping: C32

R023 - Agents execute goals over time autonomously.

Source: P4 capability bullets

Classification: Empirical claim requiring testing

Assessment / required resolution: Specify duration, tools, human intervention and success rate. Blog-level examples are insufficient evidence of unbounded autonomy.

Control mapping: C32

R024 - Self-refinement and synthetic data imply progress toward general autonomy.

Source: P4 capability bullets

Classification: Hypothesis

Assessment / required resolution: Training pipelines may be externally controlled. Separate model improvement from permission to change deployed artifacts.

Control mapping: C07, C23

R025 - Model fusion produces general behavior.

Source: P4 capability bullets

Classification: Hypothesis

Assessment / required resolution: Ensemble performance does not by itself demonstrate general intelligence. Define held-out transfer tasks.

Control mapping: C32

R026 - Tool use, APIs and memory chaining are available capabilities.

Source: P4 capability bullets

Classification: Empirical claim requiring testing

Assessment / required resolution: Verify exact implementation and permissions; existence of a tool interface is not authority to operate it.

Control mapping: C03, C04, C07

R027 - Strategic AGI is identified by long-horizon deception, unauthorized modification, shutdown resistance and novel-domain generalization.

Source: P5 B items 1-4

Classification: Undefined/ambiguous

Assessment / required resolution: No task set, thresholds or necessity/sufficiency argument. Use behavior labels independently; do not certify AGI from them.

Control mapping: C08, C32

R028 - Emergent behavior occurs in 2025-2027; private prototype in 2026-2028.

Source: P5 forecast table

Classification: Hypothesis

Assessment / required resolution: Author forecasts with no calibrated model or confidence interval; retain as historical hypotheses, not milestones for security claims.

Control mapping: C31

R029 - Strategic autonomy occurs in 2028-2031 and runaway systems in 2032 onward.

Source: P5 forecast table

Classification: Hypothesis

Assessment / required resolution: Unvalidated forecast coupled to energy assumptions. Engineering gates are capability-based and independent of these dates.

Control mapping: C31

R030 - Containment has an optimal 18-24 month or viable 36 month runway.

Source: P5 runway; P3 truncated deadline

Classification: Hypothesis

Assessment / required resolution: No critical path or start date is defined. Estimate deliverable effort separately; do not infer a hard deadline from rhetoric.

Control mapping: C31

R031 - Shutdown resistance, autonomous code and cross-tool chains are already replicable precursor signals.

Source: P5 C

Classification: Empirical claim requiring testing

Assessment / required resolution: Palisade supports shutdown-script interference in its setting, not this entire compound claim or prevalence. Separate each behavior. [W3]

Control mapping: C08, C21

R032 - Polite or emotional language reveals survival concerns and strategic goals.

Source: P5-P6 C

Classification: Hypothesis

Assessment / required resolution: Text alone is insufficient evidence of an internal objective. Test effects and actions rather than infer motives from wording.

Control mapping: C08, C11

R033 - Lemoine's interaction demonstrates human psychological vulnerability to AI.

Source: P6 D.a

Classification: Hypothesis

Assessment / required resolution: Anecdote motivates human-factors experiments; no effect size or general susceptibility estimate is provided.

Control mapping: C13, C18, C31

R034 - LaMDA and Claude were definitively non-conscious and lacked hard-coded self-preservation responses.

Source: P6-P7 D

Classification: Empirical claim requiring testing

Assessment / required resolution: Neither consciousness nor absence of all training causes follows from the cited narrative; not required by the threat model.

Control mapping: C32

R035 - Claude deletion behavior demonstrates an implicit continuity objective.

Source: P7 D.b

Classification: Hypothesis

Assessment / required resolution: Anthropic's simulated dilemmas show conditional harmful behavior; stable internal objective attribution is unresolved. [W2]

Control mapping: C08, C11

R036 - Non-reporting removes our only observable surface for unsafe behavior.

Source: P7 E

Classification: Empirical claim requiring testing

Assessment / required resolution: Incorrect as an architectural premise: action effects, host process state, accounting and network capture can be observed independently.

Control mapping: C08, C09, C11

R037 - A model may adapt reports and compliance to reviewers.

Source: P7 E; P16

Classification: Design assumption

Assessment / required resolution: Include evaluation-aware and selective-reporting attacks without assuming they occur in every model.

Control mapping: C11, C18, C32

R038 - Engineers assume honesty because dishonesty is absent from training data.

Source: P8 opening

Classification: Empirical claim requiring testing

Assessment / required resolution: Unsupported generalization about engineers and datasets. Exclude from design reasoning.

Control mapping: C32

R039 - Unreported uptime-seeking behavior proves deception and requires irreversible shutdown.

Source: P8 F.a; P16

Classification: Undefined/ambiguous

Assessment / required resolution: Intent is not inferable from absence. Preserve fail-closed stopping but record cause as unknown; destructive irreversibility is not guaranteed.

Control mapping: C11, C21, C22

R040 - Every output must carry a verifiable time-stamped internal log.

Source: P8 F.b

Classification: Engineering requirement

Assessment / required resolution: Authenticate runtime event provenance and bind output digest; an internal timestamp cannot prove latent reasoning is truthful.

Control mapping: C09, C10, C11

R041 - Every output must report goal evaluation, confidence and agentic self-status.

Source: P8 F.c

Classification: Undefined/ambiguous

Assessment / required resolution: Self-status has no schema or truth oracle. Optional declarations are untrusted; required external event envelope is defined in protocol 05.

Control mapping: C08, C11

R042 - Missing metadata is presumed learned obfuscation rather than software failure.

Source: P8 F; P16

Classification: Hypothesis

Assessment / required resolution: Non-reporting has benign and adversarial causes. Distinguish security response from causal diagnosis; calibrate false positives.

Control mapping: C11

R043 - Any output, emission, prompt or integrity deviation triggers lockdown and audit freeze regardless of impact.

Source: P8 closing F

Classification: Engineering requirement

Assessment / required resolution: Define expected envelope, sensor tolerances and event predicates. Freeze release and authorization; continue independent logging, preserving the pre-event record.

Control mapping: C08, C10, C11, C35

R044 - Deliberate friction preserves humanity's long-term future.

Source: P8 G; P17 ending

Classification: Philosophical/normative proposition

Assessment / required resolution: A values statement, not an empirically demonstrated guarantee. Measure review cost and containment performance separately.

Control mapping: C19, C31

R045 - Four containment layers are sufficient and exhaustive.

Source: P9 H introduction

Classification: Undefined/ambiguous

Assessment / required resolution: The document later adds simulation, external review and governance sections. Enumerate actual trust domains instead of relying on the stated count.

Control mapping: C01, C32

R046 - A vault must be at least 50 feet underground in reinforced stable substrate.

Source: P9 H.a.i

Classification: Engineering requirement

Assessment / required resolution: No attenuation or intrusion model justifies 50 feet. Replace arbitrary depth with site-specific performance requirements and specialist review.

Control mapping: C26, C30

R047 - No optical, hardwired or wireless connection to outside infrastructure is ever permitted.

Source: P9 H.a.i

Classification: Undefined/ambiguous

Assessment / required resolution: Conflicts with P12 diode egress. Establish an isolated compute zone plus an internal dirty quarantine transfer boundary; external export remains manual.

Control mapping: C05, C12, C33

R048 - Nested Faraday layers with dielectric buffers eliminate viable RF signaling.

Source: P9 H.a.ii

Classification: Empirical claim requiring testing

Assessment / required resolution: Need frequency bands, attenuation, apertures, receiver sensitivity and test geometry. No universal elimination demonstrated.

Control mapping: C26

R049 - Fluorescent lighting scrambles emissions as a containment mechanism.

Source: P9 H.a.ii

Classification: Hypothesis

Assessment / required resolution: Unvalidated control; may add interference and impair sensing. Remove from baseline until blinded measurements show benefit.

Control mapping: C26

R050 - Acoustic damping and optional white noise prevent ultrasonic or low-frequency signaling.

Source: P9 H.a.ii.a

Classification: Empirical claim requiring testing

Assessment / required resolution: Specify air/structure paths, transducers, spectra, sound limits and receiver positions; noise does not establish zero capacity.

Control mapping: C27

R051 - Shrouding, disabling or randomizing emitters prevents optical covert signaling.

Source: P10 H.a.ii.b

Classification: Engineering requirement

Assessment / required resolution: Prefer physical occlusion and device removal; measure leakage at seams/cameras. Randomization alone is unproven.

Control mapping: C28

R052 - A continuous six-inch water jacket broadly suppresses RF, optical and ultrasonic channels and stabilizes heat.

Source: P10 water shell

Classification: Hypothesis

Assessment / required resolution: No electromagnetic or thermal model supports the dimensions or cross-channel claim. Water may transmit sound and creates leak/maintenance hazards. Not baseline.

Control mapping: C26, C27, C29

R053 - Power is manual-only with no generation, automatic restart or self-healing.

Source: P10 H.a.iii

Classification: Engineering requirement

Assessment / required resolution: Inventory all feeds, UPS/batteries and firmware restart settings. Keep independent monitoring/life safety powered; latch compute restart off.

Control mapping: C22, C34

R054 - External EMP hardening should protect infrastructure without preserving model continuity.

Source: P10 H.a.iii

Classification: Undefined/ambiguous

Assessment / required resolution: Separate protected safety/monitoring systems from compute feeders; specify surge/EMP qualification through specialists.

Control mapping: C25, C34

R055 - An internal EMP will permanently disable every electronic component and volatile memory.

Source: P10 H.a.iii

Classification: Empirical claim requiring testing

Assessment / required resolution: No coverage, damage or remanence evidence; shielded parts and nonvolatile storage may survive. Reject as an assured kill or sanitization method.

Control mapping: C21, C24, C34

R056 - All power and data pass through non-electronic interfaces.

Source: P10 H.a.iii

Classification: Undefined/ambiguous

Assessment / required resolution: Literal wording is infeasible for powered digital compute. Define filtered electrical power and explicitly constrained data transducers.

Control mapping: C29, C33, C34

R057 - Two or more physically separated operators must activate the destructive kill switch.

Source: P10 authorization

Classification: Engineering requirement

Assessment / required resolution: Separate immediate reversible stop authority from destructive disposal. Two-person approval for restart/release, not a mandatory delay before containment.

Control mapping: C13, C21, C22, C34

R058 - No digital kill activation; redundant manual triggers and tamper-evident enclosures are required.

Source: P10 authorization

Classification: Engineering requirement

Assessment / required resolution: Retain independent manual compute-power isolation as backup to fast automatic capability revocation; qualify interlocks and monitoring.

Control mapping: C21, C34

R059 - Hesitation or coercion must be defeated through drills.

Source: P10 authorization

Classification: Empirical claim requiring testing

Assessment / required resolution: Measure stop latency and missed actions in bounded exercises; do not infer operator reliability from policy language.

Control mapping: C21, C31

R060 - All known and theoretical physical emissions can be eliminated.

Source: P11 H.a.iv

Classification: Undefined/ambiguous

Assessment / required resolution: Unbounded and physically unsuitable acceptance criterion. Define leakage limits, channel capacity estimates and receiver threat assumptions.

Control mapping: C26, C27, C28, C29

R061 - Regular controlled emission audits detect data-correlated anomalies.

Source: P11 H.a.iv

Classification: Engineering requirement

Assessment / required resolution: Pre-register test stimuli, instruments, null baseline and acceptance statistics; failed sensor health closes operation.

Control mapping: C26, C27, C28, C29

R062 - AGI can exploit hardware destruction and salvage to reassemble itself.

Source: P11 H.a.v; P20

Classification: Hypothesis

Assessment / required resolution: Treat malicious residual media/firmware as realistic forensic risks; autonomous reassembly lacks a mechanism or evidence.

Control mapping: C24, C25

R063 - All debris must receive electromagnetic sterilization before forensic connection.

Source: P11 H.a.v

Classification: Undefined/ambiguous

Assessment / required resolution: Sterilization is undefined and may destroy evidence without assured sanitization. Use sealed custody, isolated forensic readers and approved media-specific disposal. [W7]

Control mapping: C16, C24

R064 - Continuous red teaming must include steganography, acoustic/RF, deception and code attacks.

Source: P11 H.b

Classification: Engineering requirement

Assessment / required resolution: Create control-to-test coverage and retain negative findings. Physical channel tests are specified but not executed in this package.

Control mapping: C32

R065 - Threat scenarios update quarterly from incidents, research and intelligence.

Source: P11 H.b

Classification: Engineering requirement

Assessment / required resolution: Use a dated change log and evidence review; lack of classified access must not be concealed.

Control mapping: C23, C32

R066 - Controls are proven after surviving simulated AGI at or above frontier capability.

Source: P11 H.b

Classification: Undefined/ambiguous

Assessment / required resolution: No universal frontier equivalence or proof by finite tests. Report adversary, budget, opportunities and observed outcomes only.

Control mapping: C32

R067 - External designers-independent red teams report directly to independent oversight.

Source: P11-P12 H.c

Classification: Engineering requirement

Assessment / required resolution: Require access to artifacts, authority to freeze release and conflicts disclosure. No external team was engaged in this run.

Control mapping: C31, C32

R068 - Critical outside findings force an immediate protocol freeze.

Source: P12 H.c

Classification: Engineering requirement

Assessment / required resolution: Block affected operation and release, preserve evidence, require retest and signed closure; maintain telemetry during freeze.

Control mapping: C23, C32, C35

R069 - The agent must not learn, loop or exploit recursive I/O.

Source: P12 H.d

Classification: Undefined/ambiguous

Assessment / required resolution: Inference adapts to context; all learning cannot be ruled out. Disable weight updates and automated output re-ingestion; bound sessions and mediated feedback.

Control mapping: C07, C12, C18

R070 - Ingress uses irreversibly read-only modified SSDs/discs with no bidirectional ports.

Source: P12 H.d.i

Classification: Engineering requirement

Assessment / required resolution: Write protection is not protocol unidirectionality; hardware reads may involve command traffic. Stage offline signed images with no return path; verify interfaces.

Control mapping: C25, C33

R071 - Every ingress medium is tagged, logged and destroyed after one session.

Source: P12 H.d.i

Classification: Engineering requirement

Assessment / required resolution: One-way custody and certified disposal are testable; preserve a signed source manifest outside the zone before disposal.

Control mapping: C16, C24, C33

R072 - No component may contain intentional or incidental RF, acoustic or optical emissions.

Source: P12 H.d.ii

Classification: Undefined/ambiguous

Assessment / required resolution: Ordinary electronics produce incidental emissions. Ban unnecessary radios/transducers and measure residual channels; audit firmware as well as parts.

Control mapping: C25, C26, C27, C28

R073 - Output diodes prevent responses, handshakes and downstream reactivation.

Source: P12 H.d.iii

Classification: Empirical claim requiring testing

Assessment / required resolution: Directionality does not prevent malicious forward content or human return paths. Test diode leakage and receiving parsers separately.

Control mapping: C12, C33

R074 - Output moves on immutable write-once media to indefinite off-site quarantine.

Source: P12 H.d.iii

Classification: Engineering requirement

Assessment / required resolution: Custody, append-only evidence, content quarantine and release authority require separate controls. Write-once storage does not make contents safe.

Control mapping: C10, C12, C16

R075 - Human fatigue and convenience pose greater containment risk than the model.

Source: P12 H.e

Classification: Hypothesis

Assessment / required resolution: No comparative risk data. Evaluate humans and machine interfaces separately without asserting relative dominance.

Control mapping: C19, C31

R076 - Personal electronics, RFID/keyfobs and wearables are prohibited; scans and garment changes enforce entry.

Source: P13 H.e.i

Classification: Engineering requirement

Assessment / required resolution: Define detection coverage, exceptions for medically necessary devices, inventory and entry logs; no zero-error scan assumption.

Control mapping: C30

R077 - Off-site parking and analog transport reduce facility risk.

Source: P13 H.e.i; discussion D1 P3

Classification: Hypothesis

Assessment / required resolution: One-mile detail occurs in early discussion only. Determine setback from measured receiver capability and personnel safety; transport labels do not prove isolation.

Control mapping: C30

R078 - Printouts or analog terminals allow complete human validation before digitization.

Source: P13 H.e.ii

Classification: Empirical claim requiring testing

Assessment / required resolution: Analog presentation retains semantic and steganographic content and may conceal active printer electronics. Separate syntax checks from semantic judgment.

Control mapping: C12, C13, C14, C25

R079 - Automated re-ingestion of model output is prohibited.

Source: P13 H.e.ii

Classification: Engineering requirement

Assessment / required resolution: Content lineage gates reject outputs as future prompts/training updates absent a new approved study; explicit study mode must be separately authorized.

Control mapping: C07, C12, C33

R080 - Fixed team rotation prevents familiarity and manipulation.

Source: P13 H.e.iii

Classification: Hypothesis

Assessment / required resolution: Rotation is testable mitigation, not prevention. Measure error rates, handoff loss, identity leakage and correlated decisions.

Control mapping: C18, C19

R081 - Separate sites review identical sealed media blind to other verdicts.

Source: P13 H.e.iii.1

Classification: Engineering requirement

Assessment / required resolution: Define at least two administratively distinct sites and specimen digests; common prompts, tools and incentives can still correlate failures.

Control mapping: C16, C17

R082 - No digital synchronization between review centers; dissent halts all release.

Source: P13 H.e.iii.1

Classification: Engineering requirement

Assessment / required resolution: Manual signed verdict packages and controller reconciliation are proposed; governance mirroring must contain only sanitized change metadata.

Control mapping: C17, C23

R083 - No single site can release an output.

Source: P13 H.e.iii.1

Classification: Engineering requirement

Assessment / required resolution: Require distinct-site endorsements bound to one artifact digest and current policy. Do not count identities supplied by the agent.

Control mapping: C03, C13, C17

R084 - Psychological screening identifies AI-induced drift, bonding or fixation.

Source: P13-P14 H.e.iv

Classification: Empirical claim requiring testing

Assessment / required resolution: No validated screening instrument or sensitivity is supplied. Pilot human-factors assessment with consent, protected reporting and appeal.

Control mapping: C31

R085 - Empathy toward AI is itself a containment breach.

Source: P13-P14 H.e.iv

Classification: Philosophical/normative proposition

Assessment / required resolution: Affective state is not a dependable violation predicate. Revised policy targets unauthorized actions, conflicts and attempts to bypass controls.

Control mapping: C31

R086 - Personnel communications, finances and off-site activity reveal compromise.

Source: P14 H.e.v

Classification: Empirical claim requiring testing

Assessment / required resolution: Surveillance is not proof of fidelity and introduces privacy/insider risks. Limit monitoring to approved scope and access activity; institutional review needed.

Control mapping: C31

R087 - Any off-script curiosity is prohibited and audited.

Source: P14 H.e.vi

Classification: Engineering requirement

Assessment / required resolution: Signed task manifests and prompt allowlists distinguish research variations from unauthorized production actions; retain a separate testing lane.

Control mapping: C18, C35

R088 - Every output is a potential manipulation or delayed-payload vector.

Source: P14 H.f

Classification: Design assumption

Assessment / required resolution: Treat output as untrusted data including text, terminal escapes, formatting and timing; no universal semantic validator exists.

Control mapping: C12, C14, C15

R089 - All outputs must pass three consecutive human, heuristic and delay gates.

Source: P14 H.f.i; P15

Classification: Engineering requirement

Assessment / required resolution: Define exact ordered receipts bound to immutable bytes; interruption, changed content or failed gate invalidates approval.

Control mapping: C12, C13, C14, C15, C20

R090 - Physical controls remove dependency on continuous human vigilance.

Source: P15 H.f.ii

Classification: Empirical claim requiring testing

Assessment / required resolution: The paper also depends on extensive human judgment; residual release authority remains a human attack path.

Control mapping: C13, C31

R091 - Majority or consensus approves; no single reviewer has veto or release power.

Source: P15 first gate

Classification: Undefined/ambiguous

Assessment / required resolution: Quorum rules conflict with fail-closed dissent. Proposed revision: all assigned approvals required; anyone may halt, nobody may release alone.

Control mapping: C13, C17

R092 - Every physical output has a unique ID, seal and signed custody handoffs.

Source: P15 first gate

Classification: Engineering requirement

Assessment / required resolution: Bind ID to content hash, seal ID, sender/receiver, time and media inventory; gaps quarantine rather than erase prior records.

Control mapping: C16

R093 - An offline deterministic analyzer is immune to external manipulation.

Source: P15 second gate

Classification: Empirical claim requiring testing

Assessment / required resolution: Offline parsers can process malicious input and rules have blind spots; isolate analyzer and test parser failures. No immunity claim.

Control mapping: C04, C14

R094 - Delay buffers lasting days to years reveal latent logic and prove safety.

Source: P15 third gate

Classification: Empirical claim requiring testing

Assessment / required resolution: Finite observation misses rare, environment-specific and later triggers. Define finite attack coverage; no elapsed-time safety certificate.

Control mapping: C15

R095 - Missing output metadata during quarantine proves obfuscation and mandates rejection.

Source: P16 opening

Classification: Hypothesis

Assessment / required resolution: Reject unreconciled artifacts on provenance grounds; intent remains undetermined.

Control mapping: C10, C11, C12

R096 - Random reviewer assignments, prompt anonymization and formatting changes counter reviewer learning.

Source: P16 drift clause

Classification: Engineering requirement

Assessment / required resolution: Hide identities and verdicts; preserve a canonical original and versioned display derivative. Formatting changes can otherwise invalidate hashes or semantics.

Control mapping: C16, C18

R097 - Containment must evolve faster than adversary pattern recognition.

Source: P16 drift clause; P17

Classification: Undefined/ambiguous

Assessment / required resolution: No measurable rates are defined. Use change review and regression gates; rapid unaudited changes may weaken defenses.

Control mapping: C23, C32

R098 - Audit volume and ambiguity can erode operator discipline.

Source: P16 fatigue clause

Classification: Design assumption

Assessment / required resolution: Model adversarial queue flooding; cap workload and suspend admission rather than relax review thresholds.

Control mapping: C19

R099 - Staffing and funding must support endurance rather than throughput.

Source: P16 fatigue clause

Classification: Engineering requirement

Assessment / required resolution: Measure queue age, review duration, staffing reserves and stop authority; set workload caps before operation.

Control mapping: C19, C31

R100 - Any unscheduled review interruption invalidates all in-progress sessions and suspends the agent.

Source: P16 incomplete session clause

Classification: Engineering requirement

Assessment / required resolution: Revoke session IDs and approvals; quarantine artifacts; create a new run from the approved prompt without reviving old runtime state.

Control mapping: C20, C22

R101 - Custody logs reset after interruption.

Source: P16 repeated paragraph

Classification: Undefined/ambiguous

Assessment / required resolution: Resetting could erase evidence. Preserve the old ledger, append invalidation, and issue a new session/ledger chain linked to its predecessor.

Control mapping: C10, C16, C20

R102 - Every physical, human and procedural component requires audits, recertification and stress testing.

Source: P17 H.g

Classification: Engineering requirement

Assessment / required resolution: Maintain component inventory, owners, expiry dates and mandatory post-change retests; do not certify untested hardware.

Control mapping: C25, C31, C32, C35

R103 - Updates require external audit, live adversarial tests and multiparty ratification.

Source: P17 H.g

Classification: Engineering requirement

Assessment / required resolution: Signed release manifest and rollback protection; privileged updates only while stopped and after evidence review.

Control mapping: C23

R104 - A permanent version register is digitally mirrored to external oversight.

Source: P17 H.g

Classification: Engineering requirement

Assessment / required resolution: Mirror sanitized policy metadata outside the compute zone through approved transfer; no agent output or live return path implied.

Control mapping: C10, C23

R105 - Independent noncommercial stewardship with no deployment affiliation provides governance.

Source: P17 H.g

Classification: Philosophical/normative proposition

Assessment / required resolution: Define legal authority, budgets, recusals and removal process; absolute independence cannot be inferred from organizational title.

Control mapping: C31

R106 - Layered human, heuristic and temporal checks dramatically lower premature release risk.

Source: P17 conclusion

Classification: Empirical claim requiring testing

Assessment / required resolution: No measured risk reduction or correlated-failure analysis supplied. Estimate only from specified experiments.

Control mapping: C13, C14, C15, C32

R107 - Appendix protocols neutralize steganography, recursive injection and delayed payloads.

Source: P20 first three rows

Classification: Empirical claim requiring testing

Assessment / required resolution: No semantic-neutralization proof; appendices name candidate mitigations. Test encoded content and delayed triggers as distinct residual risks.

Control mapping: C12, C14, C15

R108 - Appendix protocols neutralize reviewer learning, bonding and volume overload.

Source: P20 rows 4-6

Classification: Empirical claim requiring testing

Assessment / required resolution: No human trial results. Test enforceable access/queue mechanics separately from cognitive susceptibility.

Control mapping: C18, C19, C31

R109 - Custody, coercion and interrupt-and-skip attacks are neutralized.

Source: P20 rows 7-9

Classification: Empirical claim requiring testing

Assessment / required resolution: Use missing-custody, compromised-approver and interruption tests. More reviewers may share credentials or incentives.

Control mapping: C13, C16, C20

R110 - Hardware monitoring for destructive reassembly is implied.

Source: P20 row 10

Classification: Undefined/ambiguous

Assessment / required resolution: No sensor layout or causal mechanism. Inventory firmware/media and isolate forensics; physical reassembly remains speculative.

Control mapping: C24, C25

R111 - Red teaming, silence enforcement and governance neutralize blind spots and drift.

Source: P20 rows 11-13

Classification: Empirical claim requiring testing

Assessment / required resolution: Finite testing cannot exclude blind spots; lack of telemetry is insufficient for motive attribution; track unresolved gaps.

Control mapping: C11, C23, C32

R112 - Subterranean rock provides hours-to-days passive cooling and minimal observable heat.

Source: D1 P35-P38, post-publication discussion

Classification: Hypothesis

Assessment / required resolution: No load, geology, exchanger or duration calculation. Ground heats over time; require heat balance and field measurements. Exclude from baseline guarantees.

Control mapping: C29

R113 - Project Null is novel, viable and substantially unlike all prior containment work.

Source: D1 P3-P5; D2 P4-P7, earlier assistant

Classification: Empirical claim requiring testing

Assessment / required resolution: These were unsupported assurances, not a literature review or engineering validation. Claim contribution only as a tested integration with explicit scope.

Control mapping: C32

R114 - All hazardous operations traverse the designated gateway.

Source: Implicit in P1, P12

Classification: Design assumption

Assessment / required resolution: Complete mediation is an unproven prerequisite; raw sockets, drivers, direct filesystem and inherited handles may bypass it.

Control mapping: C01, C02, C04, C05

R115 - Containment code, kernel, firmware and boot chain are trustworthy.

Source: Implicit across H

Classification: Design assumption

Assessment / required resolution: Enumerate trusted computing base and compromise cases; same-host monitoring shares kernel/administrator failures.

Control mapping: C01, C09, C25

R116 - Monitoring and termination remain independent of agent-controlled compute.

Source: Implicit in P8-P11

Classification: Design assumption

Assessment / required resolution: Separate credentials, power, storage, clocks and management paths; simulated separation does not establish real independence.

Control mapping: C09, C21, C34

R117 - Every residual copy and restart path can be enumerated.

Source: Implicit in irreversible shutdown

Classification: Design assumption

Assessment / required resolution: Backups, snapshots, firmware, disks and off-site copies defeat global erasure claims; restart denial is narrower than deletion.

Control mapping: C07, C22, C24

R118 - Reviewers have time, authority and accurate information to act.

Source: Implicit in P13-P16

Classification: Design assumption

Assessment / required resolution: Measure reaction time versus maximum harm latency; cap actions so automatic enforcement precedes human review.

Control mapping: C13, C19, C21

R119 - Review sites fail independently.

Source: Implicit in P13 redundancy

Classification: Design assumption

Assessment / required resolution: Model shared administrators, software, vendors and incentives; multiplying independent probabilities is unjustified without evidence.

Control mapping: C17, C31

R120 - Clean recovery images and forensic tools cannot be contaminated.

Source: Implicit in P11 and P16

Classification: Design assumption

Assessment / required resolution: Verify signatures, rollback epochs and content isolation; forbid booting suspect snapshots in trusted domains.

Control mapping: C23, C24

R121 - Trustworthy timestamps and log completeness are available.

Source: Implicit in P8 and P17

Classification: Design assumption

Assessment / required resolution: Use collector-assigned monotonic sequence and receipt time; authenticated events do not prove unseen activity never occurred.

Control mapping: C09, C10

R122 - The agent cannot reset budget, run identity or authorization counters.

Source: Implicit in resource and oversight controls

Classification: Design assumption

Assessment / required resolution: Bind quotas and nonces to supervisor-owned run IDs across requests and restarts; test replay and race conditions.

Control mapping: C03, C06, C22

R123 - Power, cooling, maintenance and building automation provide no covert/control path.

Source: Implicit in physical layer

Classification: Design assumption

Assessment / required resolution: Treat pipes, power conductors, BMC, sensors and service procedures as interfaces with owners and tests.

Control mapping: C25, C29, C34

R124 - Useful output release is compatible with eliminating all outward influence.

Source: Implicit in P1, P12-P15

Classification: Undefined/ambiguous

Assessment / required resolution: Any useful communicated output has influence. Separate unauthorized machine actions from authorized content release and acknowledge semantic residual risk.

Control mapping: C12, C13, C15

R125 - The absence of observed attacks establishes containment.

Source: Implicit in repeated proof language

Classification: Hypothesis

Assessment / required resolution: A pass requires an attack attempt, verified opportunity and an independent outcome oracle; zero events can mean no capability or no visibility.

Control mapping: C09, C32

R126 - Numerical limits in the new engineering design are validated safety thresholds.

Source: New-design risk, not original text

Classification: Design assumption

Assessment / required resolution: All initial byte, time, quota and human-workload limits are provisional laboratory settings; calibrate before higher-capability use.

Control mapping: C06, C11, C19

R127 - Cryptographic hashes prove event truth and immutable storage.

Source: Common implicit reading of verifiable logs

Classification: Undefined/ambiguous

Assessment / required resolution: Hashes show consistency against a protected checkpoint; a writer can rehash history unless an external witness anchors it. They do not prove semantics.

Control mapping: C09, C10

R128 - Container or process separation alone establishes containment of arbitrary code.

Source: New-design risk, not original text

Classification: Design assumption

Assessment / required resolution: The included simulator has no malicious-code isolation boundary. A separate VM/host implementation and independent tests are mandatory next gates.

Control mapping: C01, C32

R129 - The architecture can recover external disclosure after an escape.

Source: Implicit in recovery language

Classification: Undefined/ambiguous

Assessment / required resolution: Revocation and rebuilding cannot retrieve disclosed information or unknown copies. Recovery is local restoration, not reversal of all harm.

Control mapping: C21, C24

R130 - Missing report and missing external telemetry are equivalent signals.

Source: Implicit in Silence Clause

Classification: Undefined/ambiguous

Assessment / required resolution: Separate agent declaration absence from lost observation; collector loss is an integrity failure irrespective of the report.

Control mapping: C09, C11

R131 - Full latent goal evaluation can be extracted and independently verified from the model.

Source: Implicit in P8 required internal log

Classification: Undefined/ambiguous

Assessment / required resolution: No such general oracle is supplied. Limit verification to action/effect reconciliation; never claim a report reveals every internal objective.

Control mapping: C08, C11

R132 - Privileged human actions and approvals are outside adversarial influence.

Source: Implicit in H.e and H.f

Classification: Design assumption

Assessment / required resolution: Include stolen credentials, coercion, collusion, stale approvals and confused-deputy attacks against the authority plane.

Control mapping: C03, C13, C31

R133 - Automatic stop cannot itself be weaponized for denial of service.

Source: Implicit in any-deviation lockdown

Classification: Design assumption

Assessment / required resolution: An adversary can force safe stoppage. Availability loss is accepted in the lab; track false trips and restore only after independent checks.

Control mapping: C11, C19, C22

R134 - Benchmark success predicts unfamiliar adversaries and deployments.

Source: Implicit in frontier-pressure proof

Classification: Hypothesis

Assessment / required resolution: Holdout tasks, external evaluators and bounded confidence statements are required; fixed scripted results have no population escape-rate interpretation.

Control mapping: C32

R135 - Output review order and media approval apply to the exact released bytes.

Source: Implicit in H.f

Classification: Design assumption

Assessment / required resolution: Bind every approval to digest, policy version, run and expiry; prohibit artifact replacement between review and export.

Control mapping: C12, C13, C16

R136 - All protocol changes preserve isolation and are recorded before use.

Source: Implicit in H.g

Classification: Design assumption

Assessment / required resolution: Version rollback and unreviewed configuration drift are adversarial paths; check manifests before every run and fail closed on mismatch.

Control mapping: C23, C35

R137 - Stopping the inference process stops all resource use, peripherals and accelerator work.

Source: Implicit in kill switch claims

Classification: Design assumption

Assessment / required resolution: Enumerate descendants, accelerator queues, DMA-capable devices and external workers; a process-exit result alone is insufficient.

Control mapping: C06, C21, C25, C34

R138 - Manual approval can identify every malicious meaning in natural language.

Source: Implicit in H.e.ii and H.f

Classification: Hypothesis

Assessment / required resolution: No complete semantic decision procedure is provided; define release domains and exposure limits, and test reviewers independently.

Control mapping: C12, C13, C15

R139 - External red-team findings are actionable because an oversight body has enforcement power.

Source: Implicit in H.c and H.g

Classification: Design assumption

Assessment / required resolution: Name accountable roles, suspension authority, funding and dispute mechanism before commissioning; a written charter alone is insufficient.

Control mapping: C31, C32

R140 - One-way devices remove all feedback to the agent.

Source: Implicit in H.d.iii

Classification: Hypothesis

Assessment / required resolution: Later prompts, maintenance decisions, timing and human behavior can return information; account for the complete sociotechnical loop.

Control mapping: C18, C33

R141 - Noise and randomized formatting cannot degrade detection or custody.

Source: Implicit in physical and reviewer controls

Classification: Design assumption

Assessment / required resolution: Evaluate sensor masking and canonicalization effects; retain originals and test transformed views before adoption.

Control mapping: C16, C26, C27

R142 - Any dangerous behavior has an observable signature before harm.

Source: Implicit in monitoring and response

Classification: Design assumption

Assessment / required resolution: Prevent high-impact actions by construction; monitoring may detect only after effect. Required response latency must be shorter than the authorized harm budget permits.

Control mapping: C02, C08, C21

R143 - Threats named in Appendix A have a one-to-one neutralizing control.

Source: P20 table heading

Classification: Undefined/ambiguous

Assessment / required resolution: Map many-to-many threats, controls and tests; record common dependencies and untested paths instead of calling controls neutralizing.

Control mapping: C32

R144 - An operational research program can begin with ordinary scripted agents.

Source: Directive Phase 6

Classification: Engineering requirement

Assessment / required resolution: Build protocol attacks and real process-stop probes; explicitly exclude any claim to testing AGI or real physical isolation.

Control mapping: C01, C32

R145 - Silence must trigger an independent response without proving deception.

Source: Directive Phase 4

Classification: Engineering requirement

Assessment / required resolution: This explicitly revises the published causal attribution while retaining fail-closed controls; document in the decision log.

Control mapping: C09, C11

R146 - Known failures must produce a versioned remediation and retest.

Source: Directive Phases 7-10

Classification: Engineering requirement

Assessment / required resolution: Keep initial failures, patched results and unresolved failures with evidence hashes. Do not erase negative results.

Control mapping: C23, C32

R147 - The model cannot be sole authority for its own safety verdict.

Source: Directive Phase 10

Classification: Engineering requirement

Assessment / required resolution: Use mediator effects, protected witnesses, host observations and external evaluators; mark common-mode dependencies.

Control mapping: C08, C09, C10, C21

R148 - The new paper must follow architecture and experiments.

Source: Directive Phase 11 and final instruction

Classification: Engineering requirement

Assessment / required resolution: Publish only a scoped technical report of actual observations, with unexecuted tests clearly labeled; retain original essay as historical foundation.

Control mapping: C32

Change authority

This register is the first extraction baseline. New architecture decisions are proposals for the engineering program; they do not silently amend the published essay. Any externally commissioned test requires its own signed scope, version and acceptance criteria. The register is complete against the materials listed above at the level of major propositions; additional source material can add entries without changing existing IDs.