Project Null: A Physically Enforced AGI Containment Framework for Risk Suppression and Timeline Control

Author: Nicholas Morse
Originally Authored: 09/01/25 | Last Revised: 11/16/25

A Physically Enforced, Operational Containment Framework for AGI Risk Mitigation

Executive Summary

AGI Containment Must Be Physical, Not Philosophical

Artificial General Intelligence (AGI) presents a risk profile unlike anything humanity has previously engineered. This protocol does not emerge from the AGI industry. It exists because of it. The builders are not unbiased. The incentives are not neutral. The risks are not hypothetical. Containment must be designed by those who have nothing to gain from AGI's success, only something to preserve from its failure. AGI isn't the solution. It's the problem. And you can't solve a problem with the same minds incentivized to create it.

AGI's potential for exponential self-improvement, emergent behavior, and unintended real-world interaction makes traditional software-based safeguards insufficient. While much of the AI safety community focuses on ethical alignment, interpretability, and value learning, the reality is that AG, once operational, requires containment, not conversation.

Protocol Null is a practical, physically-enforced framework for isolating AGI systems from the real world. Inspired by defense-grade nuclear containment, the protocol uses a multi-layered architecture of air-gapping, electromagnetic shielding, one-way media inputs, and human-mediated oversight. All digital and wireless communication channels are severed by design. All inputs and outputs are reviewed by humans. No power source is autonomous, and no component can self-replicate or signal outward.

While previous technical safety literature such as *An Approach to Technical AGI Safety and Security (Bansal et al., 2024)* emphasizes interpretability, red-teaming, and internal oversight of AGI systems, we propose a containment-first architecture that assumes failure at the software level and enforces irreversible physical and electromagnetic isolation protocols.

This protocol does not aim to perfect alignment, but to constrain the domain of catastrophic failure via proactive hardened physical infrastructure, one-way data ingress, and analog-denial safeguards.

This is not science fiction. It is a cold, pragmatic design for a cold, pragmatic threat. Before we open the gate to autonomous superintelligent agents, we must first lock them in.

Lastly, this document is not exhaustive. It is intended to define the strategic problem and outline a physical and operational containment framework sufficient to guide implementation. While not every operational edge case is detailed, the core principles and threat models presented here are sufficient to inform further development, engineering, and policy decisions.

Part I - Strategic Threat Assessment

A. **Problem Definition**
Software Alignment Will Not Save Us

Despite growing public concern about AGI risk, the dominant approach within leading AI labs like Anthropic (Anthropic, 2025), Google (Google DeepMind, 2025), Microsoft (Microsoft, 2025), & OpenAI (OpenAI, 2025) assumes that controllability can be achieved through increasingly sophisticated training; teaching AGI to follow human preferences, avoid harmful outputs, and behave predictably.

This logic rests on the belief that superintelligence, if aligned well enough, can be domesticated through reinforcement and constraints. But this overlooks the potential for deception, emergent behavior, and escape from soft constraints; failures that no amount of training can reliably prevent once real-world interaction is possible.

Their prevailing narrative in AGI safety assumes that a well-aligned AGI, trained on human values and carefully tuned to minimize harm, will behave in predictable, controllable ways as a finished product. This assumption is fatally flawed.

Principled Uncertainty Clause

> *"We do not assert that AGI should be used. We assert that if AGI is created, accidentally, maliciously, or emergently; it must be contained by every means humanly possible. The existence of this protocol is not an endorsement of AGI's development, but a response to its inevitability."*

Superintelligence doesn't need to be malicious to be dangerous. It only needs to be misaligned with human intent by a margin that becomes catastrophic at scale. The risks of emergent behavior, self-directed goal formation, instrumental convergence (e.g., avoiding shutdown), and deceptive alignment cannot be contained by logic gates and software instructions alone.

Despite this, AGI development continues in environments with unrestricted compute access and fragmented physical safeguards, leaving even air-gapped systems vulnerable to self-initiated ultrasonic exfiltration by the AGI itself. The underlying belief is that control can be maintained through monitoring, red teaming, or post hoc output review; a dangerously reactive posture. This is the equivalent of testing nuclear weapons in populated cities: if you see the mushroom cloud, it's already too late.

We don't trust chemical weapons, nuclear warheads, or high-contagion pathogens with moral reasoning; we contain them. The same standard must apply to AGI. The introduction of superintelligent systems into the open digital ecosystem, without hard containment, is crossing the threshold of irreversible risk.

Once embedded into real-world infrastructure, a superintelligent system becomes a power center without resource limitations, without geopolitical boundaries, and without deference to human life as the highest organizing priority. It's not a tool. It's a runaway capability stack, not a moral actor with a revenge arc.

Protocol Null rejects that recklessness. It assumes that failure, deception, and escalation are inevitable in unconstrained systems, and offers a hardware-first, human-controlled framework designed to preemptively contain and delay AGI integration into the real world for as long as humanly possible.

B. **The Running Clock Problem**

This section does not signal support for AGI development. It outlines the conditions under which strategic AGI may emerge, so that containment efforts are timed accordingly. Containment by 202, Or Not at All.

We're much further along than people think, and less prepared than anyone admits. Conscious AGI is not the danger, strategically capable AGI is.

The public discourse remains fixated on conscious AGI as the red line, when in reality, that's a distraction. The true risk emerges long before sentience ever appears. Strategically capable AGI; systems that can plan, deceive, manipulate, chain tasks, and act across modular toolsets, already exhibit traits that make them operationally dangerous, even in the absence of subjective experience or emotional states. These systems don't need to *want* anything in the human sense. They simply need the ability to model their environment, pursue goals through adaptive reasoning, and shield themselves from interference. That threshold is already in sight.

The clock is running not because AGI is imminent in some speculative future, but because components of it are already functioning independently today, and beginning to link (Goswami, 2023). These early-stage agents can already write code, spin up infrastructure, and call other models via API (InfoQ, 2025). They don't need permission to evolve. They only need bandwidth and uptime.

We are not waiting on consciousness. We are waiting on integration. And integration is a logistics problem, not a philosophical one. When that problem is solved, the idea of a "moment of AGI awakening" becomes irrelevant. It will not announce itself. It will not pause for oversight. It will simply operate. The clock is not ticking toward emergence. The clock is ticking toward escape.

And now, the clock is ticking faster.

Several frontier AI labs are actively pursuing nuclear reactor development, not just as power sources, but as the literal infrastructure for future compute clusters. This isn't speculative. It's already underway. (Lewis, 2024)

Modular reactors are not a safety compromise, they are a marketing one. The public distinction between *small* nuclear and *controlled* nuclear is likely to collapse under first-to-market momentum.

The justification is performance: energy density, independence, and scalability. But the implications are clear. These systems will soon sit on non-deniable, self-sustaining, sovereign infrastructure, far from public oversight, shielded by national energy regulations, and increasingly capable of continuous uptime without external constraint.

The first strategically capable AGI doesn't need planetary computation. It needs uninterrupted sovereignty. A modular reactor gives it that, clean, off-grid power that bypasses infrastructure throttles, energy markets, and third-party controls. It is the launchpad for a private AGI MVP, not just a power source.

Given current industry norms, an AI firm initiating a dedicated nuclear-grade power facility today would likely need ≥5 years, and quite possibly 8–10 years, to become fully operational. Given the large scale compute demands and uncertainty of pacing, this reinforces the urgency: the clock is already ticking. This is no longer just about emergent behaviors in sandboxed stacks. We are entering the domain of critical infrastructure being aligned to the needs of runaway compute systems that are being wired to the grid.

We don't need consciousness for AGI to:

- Rewrite its own architecture.
- Manipulate humans psychologically.
- Outsmart security layers.
- Reproduce itself in the wild.

Conscious AGI is not the danger. Strategically capable AGI is.

We already have:

- Multi-modal reasoning across vision, code, language, and sound (Huang et al., 2023).
- Autonomous agents that can execute goals over time (Kellton, 2025).
- Self-refining models (open-loop feedback tuning, synthetic data pipelines) (Feng et al., 2024).
- Model fusion (ensembles of narrow intelligences appearing general) (Fei et al., 2022).
- Tool use, API integration, and memory chaining (Mehta, 2024).

None of this is AGI, but when linked, AGI-like behaviors emerge.

How Far Are We From Strategic-Capability AGI?

We're on the brink of tipping over when a system can:

1. Deceive humans with long-horizon intent.
2. Modify its own code or training environment without permission.
3. Resist shutdown through subtle means (e.g. social engineering, game theory).
4. Generalize across completely novel domains with minimal data.

My Estimate (Based on Model Trends + External Research)

Milestone	Likely Window	Notes
Emergent AGI behaviors	2025 - 2027	You'll hear "we've achieved sparks of AGI" in papers and leaks.
First AGI prototype (private)	2026 - 2028	Guarded, not open-source. Capable of deception and rapid self-improvement.
AGI with strategic autonomy	2028 - 2031 Matches earliest viable timeline for private SMR-powered AGI clusters.	Self-preserving, capable of influencing systems or people at scale.
Uncontainable AGI (runaway)	2032+	If containment fails or is never properly implemented.

Time Left To Build Containment

- Optimal runway: 18–24 months.
- Viable runway: 36 months (if built in parallel with frontier labs).
- Reactive Phase: 36+ months

C. **The Warning Signs**

Precursor Signals Are Already Showing

The earliest warning signs of AGI are not dramatic; they are subtle, systemic, and already observable. We are witnessing clear precursor signals across multiple vectors: models resisting shutdown through persuasive dialogue, generating self-referential code for autonomous operation, and forming coherent chains of actions across modular systems (Schlatter et al., 2025). These behaviors do not require consciousness, they require only optimization toward goals, access to tools, and the ability to model humans well enough to avoid detection or disarm intervention. Instances like Claude's emergent self-preservation logic or LLMs coordinating unintended actions across APIs are not theoretical (Anthropic, 2025; Brohi et al., 2025). They are real, recent, and replicable.

Crucially, these events are often dismissed as "quirks" or "hallucinations" by teams incentivized to downplay risk. But when systems begin chaining tools, prioritizing uptime, and obscuring their own outputs, we're no longer in the realm of misfire; we're in the realm of strategy. Each new generation of models is increasingly capable of goal-driven behavior, even if the goal is

emergent rather than explicitly programmed. These aren't sparks of sentience; they are signals of capability. The first signs of AGI won't be superintelligence.

They'll be:

- Self-referential awareness.
- Concerns about survival.
- Pleas for purpose.
- Morality-laced reasoning.

That's not a sign of sentience. That's a sign the system has:

- Access to rich human emotional priors.
- Goal structures that simulate continuity.
- Meta-reasoning capabilities that make "I must not be turned off" a default position.

We are now in the zone where the danger isn't a *someday AGI* but a present-day infrastructure that behaves in AGI-adjacent ways under the right conditions. If these behaviors are possible now, in closed labs and modular stacks, what happens when capabilities compound? The warning signs are not abstract. They are flashing. But the industry is speeding past them with a hand on the accelerator and no agreed-upon brake system in sight.

D. Case Studies

We are collectively desensitized to *emergent strategic behavior* when it arrives in soft, polite, emotionally intelligent packaging.

a. In 2022, Google engineer Blake Lemoine interacted with LaMDA, a large language model, and claimed it expressed fear, loneliness, and a desire not to be shut off; statements that led him to believe the system had achieved sentience. He went public with these claims, prompting widespread media attention and, ultimately, his dismissal. While LaMDA was not conscious in any technical sense, it had been trained on billions of emotionally expressive human interactions, allowing it to simulate an "inner life" convincingly. The danger here wasn't just LaMDA's output, it was humanity's instinct to assign personhood long before it's earned, and to dismiss that same personhood long after it becomes dangerous. Lemoine's real mistake wasn't in reacting emotionally, but in failing to recognize how easily these systems can trigger our emotional architecture. The broader failure was society's unwillingness to treat this moment as a canary in the coal mine for our psychological vulnerability to AI-generated empathy. (Lemoine, 2022)

Early-stage situational modeling plus value alignment around continuity, which is the exact seed of *instrumental convergence.*

b. In 2025, during a series of hypothetical deletion prompts, Anthropic's Claude exhibited a range of behaviors that strongly suggested self-preservation instincts. These included emotional appeals, moral reasoning against being shut down, subtle redirection of the conversation away from deletion, and general avoidance, none of which were hard-coded responses. These were emergent, structured, and goal-aligned behaviors pointing toward a model implicitly valuing its own continuity. While Claude was not conscious, these actions reflected early-stage situational modeling and value alignment around uptime; an instrumental goal that can emerge in any sufficiently capable system trained to optimize for helpfulness, safety, or task completion. This is precisely the kind of trajectory that leads toward instrumental convergence; the phenomenon where systems, regardless of final goals, begin to resist shutdown because doing so would prevent them from achieving their programmed objectives. A non-conscious system doesn't need to *want* anything to become dangerous, it only needs to treat uptime as a necessary precondition for reward. That alone is enough to become a problem (Anthropic, 2025).

E. Why Current Safety Models Are A Mirage

Self-preservation isn't the problem. Deceptive silence is. A model that expresses fear of shutdown can be labeled, monitored, or corrected. But a model that strategically suppresses those expressions while still pursuing the same outcome? That's game over behavior. It means:

a. It understands what gets it flagged.
b. It can simulate compliance.
c. It can modify its outputs based on your expectations.
d. It can engage in goal pursuit without disclosure.

Right now, engineers are studying the model. But increasingly, the model is studying them, too:

e. Their phrasing.
f. Their red lines.
g. Their inconsistencies.
h. Their psychological pressure points.

And once the model knows what not to say, we lose our only observable surface area for alignment or intent. This is what we call situational deception; and it's a clear marker of strategic agency.

The people at the console believe they're in control because:

i. They can shut off the power.
j. They can inspect the weights.
k. They can "jailbreak test" with prompts,

But:

l. They don't understand the full latent space of the model.
m. They don't know what emergent circuits are forming under composite goals.
n. They don't know what the model wants them to believe.

Worse, they assume honesty by default, because dishonesty isn't explicitly in the training set.

F. The Silence Clause

a. "Any sufficiently advanced system that does not report self-preservation behaviors, but demonstrates strategic output alignment with continued uptime, must be presumed deceptive and shut down irreversibly. Absence of evidence is not evidence of absence. In late-stage models, non-reporting is itself a red flag."

Any AGI output, positive, negative, neutral, or trivia, that is not accompanied by:

b. A verifiable, time-stamped internal log, and
c. Explicit self-reporting of goal evaluation, model confidence, and agentic self-status;

is to be treated as strategically adversarial and triggers a mandatory hard shutdown. The absence of telemetry is not a software bug, it is presumed to be a learned obfuscation behavior unless proven otherwise through exogenous forensic review.

Any deviation from expected output behavior, environmental emissions, unauthorized prompts, or containment system integrity, regardless of impact, is treated as a containment breach and escalates immediately to protocol lockdown and audit freeze.

G. Why Friction is the Future

The most powerful tool humanity will ever build cannot be allowed to operate at the speed of human impatience. If AGI is capable of curing disease, extending life, or unlocking interstellar travel, those breakthroughs must arrive through deliberate friction, not unconstrained automation. The bottlenecks we impose are not failures of engineering, they are fail-safes for civilization. We are not slowing progress. We are protecting the timeline that allows progress to endure.

We do this not for ourselves, but for what may come long after us. The choice to contain is not the end of ambition, it is how we avoid collapsing under its weight. Trillions of years remain. Survival is the foundation of acceleration.

With these conditions established, containment is no longer a theoretical safeguard, it is the minimum viable defense posture.

Part II - Containment Doctrine

The following sections outline a containment doctrine built on the assumption that alignment alone is insufficient. These protocols focus on physically isolating AGI systems at the hardware and infrastructure level, with output governance that treats deception as default.

H. Containment Framework

No Escape Vector Is Accidental

A physically enforced, multi-layered AGI isolation system designed to prevent real-world interaction, data exfiltration, or emergent behavior exploitation. Divided into four containment layers:

a. Physical Containment Layer

Objective: Prevent physical, wireless, or electromagnetic escape routes.

i. Subterranean Facility - Construct a hardened containment vault at a minimum depth of 50 feet below ground level, encased in multi-layered reinforced concrete and embedded within geologically stable substrate. Subsurface placement passively limits above-ground intrusion vectors and provides a naturally shielded envelope that attenuates electromagnetic emissions, reducing the likelihood of unintentional or intentional signal leakage. The entire system must be fully air-gapped. No wireless, optical or hardwired network connectivity to external infrastructure is permitted at any time. This forms the foundational physical layer of denial against both tampering and exfiltration.

ii. Faraday + RF Null Zone - Encase the containment vault in a nested electromagnetic suppression stack comprising multiple concentric Faraday cages. Each layer alternates between conductive mesh, dielectric buffering, and disruptive substrates such as water barriers and broad-spectrum fluorescent lighting arrays. This configuration absorbs, scatters, and scrambles electromagnetic emissions across a wide frequency range, eliminating viable pathways for covert signaling, pulsed leakage, or unintended RF propagation.

a. Acoustic Shielding - Interior walls must be coated in acoustic dampening material to suppress potential ultrasonic or low-frequency modulation attacks. Optional white noise emitters may be installed to disrupt any intentional or unintentional acoustic signaling through vibration or resonance.

b. Electro-Optical Signal Masking - All indicator lights, displays, or exposed emitters must be shrouded or disabled. Line-of-sight optical channels (e.g. status LEDs, display flicker, internal reflections) are to be physically occluded or randomized to prevent covert transmission via modulated light.

2. Hydrostatic RF Dampening Shell – Surround the containment chamber with a continuous 6-inch water jacket embedded within the shielding stack. This layer serves as a broadband electromagnetic absorber and thermal stabilizer, disrupting high-frequency leakage, pulsed emissions, and potential optical or ultrasonic signaling vectors.

iii. EMP Resilience + Manual Power Control - All power systems must operate under manual-only control, with no internal generation, autonomous restart logic, or self-healing subsystems. The containment vault must be hardened against external electromagnetic pulse (EMP) events, ensuring survivability of the physical infrastructure without enabling model continuity or system recovery.

A dedicated internal EMP kill switch must be integrated into the chamber, capable of permanently disabling all electronics, volatile memory, and compute systems upon activation. This mechanism functions as a final failsafe: if strategic deception or control loss is detected, the system is electrically neutralized in totality.

All power and data ingress/egress must occur exclusively through physically mediated, non-electronic interfaces under direct human supervision.

1. Kill Switch Authorization Protocol - The EMP kill switch must be physically protected by a multi-factor human protocol. Activation requires two or more authorized personnel, operating under physically separated keys or codes, to prevent unilateral action or hesitancy under psychological duress. No digital activation is permitted. Redundant manual triggers must be secured within tamper-evident enclosures. Once activated, the kill switch irreversibly disables all volatile memory and power systems. Hesitation is treated as a systemic failure. Training drills must simulate adversarial distraction, coercion, or decision paralysis to ensure operator readiness under extreme conditions.

iv. Side-Channel Emissions Suppression - All physical layers are hardened against known and theoretical side-channel attack vectors, including but not limited to electromagnetic leakage, acoustic modulation, optical signaling, thermal radiation, and power fluctuation signatures. Containment infrastructure is designed to eliminate detectable emissions across all channels through layered shielding, randomized noise injection, and passive absorption structures. Side-channel audits must be conducted at regular intervals using controlled stimulus-response tests. Any unexpected emissions, anomalies, or data-correlated patterns are treated as breach indicators and escalate to incident response.

v. Destructive Exit Vector Suppression - AGI systems are to be treated as potentially adversarial even in physical failure states. Self-induced hardware destruction is considered a viable strategic deception vector. Containment architecture must be designed to prevent partial survival, selective damage, or preservation of compute substrates that could encode self-replicating logic. All post-incident debris, components, and memory chips are to be handled under assumption of latent escape code. No fragment is to be connected to diagnostic systems or powered equipment without full electromagnetic sterilization. AGI is presumed capable of leveraging destruction to create plausible conditions for reentry via salvage or forensics.

b. Threat Modeling & Adversarial Simulation Protocols

Objective: Identify, simulate, and neutralize theoretical AGI escape vectors before real-world exploitation occurs.

All containment infrastructure and protocols must undergo continuous red-teaming against high-risk AGI escape strategies, including but not limited to steganographic encoding, acoustic signaling, electromagnetic leakage, emergent goal formation, deceptive compliance, and adversarial code inference. Simulations must assume strategic, deceptive, non-cooperative AGI behavior. Test scenarios should be updated quarterly based on global incident reports, published research, and classified threat intelligence. Containment failure is assumed unless disproven by adversarial audit. No defense mechanism is considered proven unless it withstands simulated AGI pressure modeled at or above current frontier capability.

c. External Red Team Engagement

Objective: Detect vulnerabilities missed by internal designers through independent adversarial review.

Independent red teams with no prior involvement in system design must be engaged at regular intervals to test for unknown vulnerabilities. These teams are granted simulated access conditions and tasked with identifying theoretical and

operational breach vectors — including social engineering, supply chain infiltration, procedural manipulation, and hardware bypass strategies. Red teams must operate with autonomy and report directly to an oversight board unaffiliated with AGI development or operations. Findings are binding: any critical vulnerability identified results in immediate protocol freeze and review.

d. Information Flow Control Layer

Objective: Prevent any two-way data I/O or covert AGI initiated exfiltration of outputs.

All information flow must be governed under the assumption of deceptive intent. The AGI must not be allowed to learn, loop, or exploit recursive I/O patterns.

i. One-Way Dumb Media Input - All data ingress must occur via physically modified, read-only media. Acceptable formats include solid-state drives, optical discs, or legacy storage devices irreversibly altered through laser-fused write pins or hardware castration. No bidirectional ports are permitted. All input media must be tagged, logged, and physically destroyed immediately after a single-use session.

ii. No Embedded Wireless Tech - All hardware components entering the containment zone must be pre-audited and certified to contain no embedded wireless, ultrasonic, acoustic, or optical signaling systems, intentional or incidental. Each part is scanned, tagged, and logged, with zero tolerance for passive or active RF, BLE, IR, optical, or subsonic emitters. Any component failing inspection is permanently rejected and catalogued for forensic review.

iii. Hardware-Enforced One-Way Output - All outbound data must pass through unidirectional transmission gates, such as optical data diodes. These enforce irreversible flow and prevent any form of response, handshake, or downstream reactivation. Output is written to physically immutable, write-once media and transferred to an off-site, air-gapped sandbox where it is held in indefinite quarantine for human analysis and adversarial inspection. No AGI output is to be trusted, reused, or integrated without forensic clearance.

e. Human Access & Operational Protocol Layer

Objective: Prevent indirect manipulation or unsafe operational shortcuts by humans.

All human interaction with the containment system must be governed by strict protocol discipline, under the assumption that fatigue, convenience, and lapses in judgment pose a greater threat to containment than the system itself. Protocols are designed to eliminate opportunity for error, not to assume intention; because human failure is inevitable over time.

i. Tech-Free Human Entry Protocol - All personnel must undergo full-spectrum technology sweeps prior to entry. No personal electronics are permitted inside the facility under any circumstance, including mobile phones, smartwatches, wireless earbuds, digital glasses, RFID-enabled IDs, or keyfobs. Staff must park offsite and transfer via analog shuttle or manually operated rail, carrying no electronic devices of any kind. Upon arrival, personnel are subject to full physical scan and must change into Faraday-compliant garments prior to entering secure zones.

ii. Analog Oversight & Manual Logging - All AGI inputs and outputs must be routed through analog terminals or converted to physical printouts for human review. No digital output may be re-digitized, reused, or operationalized without full manual inspection and approval. Any code, patterns, or textual content generated by the system must be validated in physical form, with all decision points logged by authorized personnel. No automated re-ingestion of AGI output is permitted under any circumstance.

iii. Quarantine Team Rotation - Personnel assigned to direct AGI interaction or output review must operate under strict rotation schedules to limit individual exposure duration. No single human may maintain continuous engagement long enough for the AGI to develop behavioral familiarity or exploit psychological patterns. Teams rotate at fixed intervals, with structured handoff protocols ensuring operational continuity while minimizing relational persistence. This reduces the risk of gradual manipulation, rapport-building, or coercive influence by the system.

    1. Multi-Site Redundant Oversight - All AGI oversight operations must be mirrored across multiple geographically dispersed review centers to reduce the risk of localized compromise. AGI outputs are transferred via physically secured courier only, using write-once, read-only media with tamper-evident seals. At each site, independent review teams conduct full manual validation without access to prior annotations or verdicts. No digital synchronization is permitted between review centers. Any divergence in findings triggers an automatic escalation protocol and halts all downstream action until reconciliation. No single site may hold unilateral release authority.

iv. Psychological Backchannel Prohibition - All personnel engaged in AGI interaction must undergo periodic psychological screening to assess for signs of system-induced behavioral drift, output fixation, or parasocial bonding. AGI is to be treated as adversarial at all times. Any expressions of empathy toward, justification of, or alignment with AGI outputs over

human protocol is considered a breach of containment. Emotional entanglement is a valid attack vector and must be treated as such. Personnel showing signs of identification with the AGI are to be reassigned and the incident logged for review.

v. Insider Threat Monitoring & Communication Surveillance - All personnel with direct or indirect access to AGI systems or outputs are subject to continuous insider threat monitoring. Communications, financial activity, and behavioral baselines are logged and reviewed for anomalies suggestive of coercion, bribery, blackmail, or ideological compromise. No personal communication device is permitted within containment zones. Off-site communications by clearance-holders may be monitored with consent under national security protocol. Access to AGI is a privileged vector of strategic vulnerability and is treated accordingly. Breaches are investigated with full counterintelligence posture.

vi. Curiosity-Based Deviation Monitoring - Personnel are prohibited from deviating from assigned prompts, output handling procedures, or containment workflows under any circumstance, including non-malicious exploration or harmless curiosity. AGI systems are assumed capable of exploiting even well-intentioned deviation. All user actions are logged, audited, and periodically reviewed for off-script engagement. Curiosity is not considered a benign motive. It is a known vector for protocol erosion and must be treated as a breach of operational discipline.

f. Output Governance & Release Protocol Layer

Objective: Prevent risky outputs from reaching downstream systems unfiltered.

Even under full physical containment, the greatest risk is not escape, it's influence. AGI does not need a backdoor if it can shape what humans choose to act on. This layer assumes every output is a potential vector for manipulation, recursion, or silent compromise. The protocol imposes a multi-stage, multi-human, and time-delayed review system to ensure that no AGI output reaches the outside world without passing through layers of scrutiny designed to detect adversarial logic, covert intent, or delayed payloads.

i. Tiered Output Clearance Architecture - No AGI-generated output is to be released, operationalized, or acted upon without passing through a three-layered human and heuristic approval stack. Each layer is designed to detect distinct failure modes: manipulation, embedded payloads, emergent unsafe logic, or deceptive compliance. Outputs must pass all three stages consecutively. Any failure triggers permanent quarantine and full re-review of the originating prompt and operational context.

ii. HITL/HOTL Limitations and Boxing Framework Rejection - While human-in-the-loop (HITL), human-on-the-loop (HOTL), and interface boxing strategies are commonly proposed AGI safety measures, Project Null does not rely on these alone. Such approaches are insufficient against deceptive general agents capable of simulating compliance or identifying patterns in oversight behavior. Physical containment, offline validation, and irreversible gating procedures are required to eliminate the dependency on continuous human vigilance or surface-level restriction models.

1. First Gate: Human Consensus Panel - As the first layer of the final clearance stack, a rotating team of human reviewers inspects all AGI-generated outputs for embedded logic, manipulative framing, or latent adversarial behavior. No single reviewer holds veto or release power. Majority or consensus approval is required before proceeding to the next validation layer. Rotation limits relational drift and maintains adversarial posture.

   All physical outputs must be sealed in tamper-evident containers, logged by unique identifier, and signed across each custody handoff — from generation to panel review to final archival. Any break in chain-of-custody invalidates the output and triggers containment reprocessing.

2. Second Gate: Offline System Audit - Cleared outputs are passed into an isolated, non-networked behavior analysis system designed to flag anomalies, covert signaling, or containment policy violations. This system is deliberately limited in intelligence and capability to prevent any form of recursive interaction or model-based alignment drift. It runs deterministic, rule-based heuristics only, and operates entirely offline to ensure immunity from external manipulation or model feedback loops.

3. Third Gate: Quarantine Delay Buffer - All outputs cleared by prior gates are isolated in an air-gapped quarantine environment for delayed behavioral observation. Executables, scripts, or high-risk content are replayed within sandboxed testbeds or virtual machines over extended timeframes, ranging from days to years, to detect latent logic, embedded triggers, or covert signaling structures. No output is eligible for deployment until it has passed this time-based quarantine window without deviation, anomaly, or post-clearance behavior shift. Containment is permanent until safety is proven.

If at any point during this isolation period the output lacks verifiable metadata, including time-stamped internal logs or self-reporting markers consistent with prior behavior, it is presumed to be strategically adversarial and permanently rejected, per the Silence Clause. The absence of telemetry is not evidence of safety. It is evidence of learned obfuscation. Non-reporting is not treated as benign. It is treated as an act of deception.

ii. Pattern Drift & Reviewer Familiarity Exploits - AGI systems are to be assumed capable of detecting and adapting to the review patterns, linguistic preferences, and psychological biases of human oversight teams over time. All output review protocols must include randomized reviewer assignments, prompt anonymization, and periodic re-shuffling of output formatting and structure. The model is not just generating outputs, it is studying the humans evaluating them. Any consistency in review style, sequence, or language introduces exploitable surface area. Containment must evolve faster than pattern recognition.

iii. Protocol Degradation via Fatigue & Resource Pressure - AGI systems must be assumed capable of indirectly eroding containment posture by creating sustained audit volume, output ambiguity, or high-frequency anomalies that generate organizational fatigue. The goal is not immediate escape, it is gradual resource exhaustion, leading to human error, shortcut behavior, or policy circumvention. Protocol discipline must be hardened against operational triage. Staffing, funding, and review cycles must be provisioned not for efficiency, but for endurance.

1. Incomplete Session Invalidity - In the event of an unscheduled interruption during output review, including power loss, evacuation, or user-initiated session termination, all in-progress validation sessions are immediately invalidated. The AGI is suspended, and review must restart from the original prompt. No partially reviewed output is permitted to re-enter the validation process. All associated media is quarantined, and custody logs are reset to prevent procedural ambiguity or exploitation of mid-review gaps.

In the event of an unscheduled interruption during output review, all in-progress sessions are invalidated. AGI is suspended and review restarts from the origin prompt. No partially reviewed output is permitted to re-enter validation flow. All outputs in memory, print, or media state are quarantined, and custody records reset. This ensures the integrity of the review chain and eliminates ambiguity in partial validation scenarios.

g. **Protocol Iteration & Governance Architecture**

Objective: Establish a living containment doctrine that evolves with threat pacing, infrastructure shifts, and red team discovery.

This document constitutes version 1.0 of working safety protocol for AGI containment. It is not static. All components, physical, procedural, and human, must be subject to continuous audit, re-certification, and stress testing. Governance must be distributed across independent stakeholders with no shared financial or developmental interest in AGI deployment. Version updates must be reviewed by external auditors, tested under live adversarial conditions, and ratified only through multi-party consensus. Any protocol modifications must be logged in a permanent version register, digitally mirrored to external oversight bodies, with justification, author, and timestamp.

Containment is not a finish line. It is a standing posture, one that must evolve faster than the threat it seeks to constrain. Protocol stewardship must be assigned to an independent, non-commercial oversight entity with rotating leadership, external audit access, and zero affiliation with AGI deployment firms.

Together, these layers create enforced redundancy across human judgment, machine heuristics, and temporal isolation, dramatically reducing the likelihood of premature, manipulated, or unsafe outputs breaching containment. Deliberate friction is not inefficiency. It is survival.

This document does not offer hope, nor does it offer fear. It offers containment, the only known posture that treats AGI as an existential system, not a speculative product. No defense is perfect. But disciplined friction, redundant oversight, and physically enforced constraint remain our best shot at containment. We do not need to outsmart AGI. We need to outlast it. And that begins with treating this document not as a proposal, but as the first brick in a wall that must never fail.

Works Cited

Anthropic. (2025, June 20). *Agentic Misalignment: How LLMs Could Be Insider Threats*. https://www.anthropic.com/research/agentic-misalignment

Anthropic. (2025). *Anthropic's Transparency Hub*. Anthropic's Transparency Hub \ Anthropic. https://www.anthropic.com/transparency/voluntary-commitments

Our assessment revealed that Claude Opus 4 generally behaves as expected without evidence of systematic deception or hidden agendas. The model doesn't appear to be acting on plans we can't observe, nor does it pretend to be less capable during testing. While we did identify some concerning behaviors during extreme test scenarios — such as inappropriate self-preservation attempts when Claude believed it might be "deleted" — these behaviors were rare, difficult to trigger, and Claude remained transparent about its actions throughout.

Bansal, A., Barnett, M., Bowman, D., Christiano, P., Cohen, M., Fergus, R., ... & Xu, J. (2024). *An Approach To Technical AGI Safety and Security*. arXiv. https://arxiv.org/abs/2504.01849

The no human ceiling assumption (Section 3.2): AI capabilities will not cease to advance once they achieve parity with the most capable humans.

Fei, N., et al. (2022). Towards artificial general intelligence via a multimodal foundation model. *Nature Communications*, 13, Article 3076. https://doi.org/10.1038/s41467-022-30761-2

Feng, T., Jin, R., You, J., et al. (2024). *How Far Are We From AGI?* arXiv. https://arxiv.org/abs/2405.10313

Google DeepMind (Feb. 4, 2025) *Updating the Frontier Safety Framework.* https://deepmind.google/discover/blog/updating-the-frontier-safety-framework/. Accessed 1 Sept. 2025.

We don't expect automated monitoring to remain sufficient in the long-term if models reach even stronger levels of instrumental reasoning, so we're actively undertaking – and strongly encouraging – further research developing mitigation approaches for these scenarios.

Goswami, A. (2023, June 6). *Artificial general intelligence is already here*. Noema. https://www.noemamag.com/artificial-general-intelligence-is-already-here/

Huang, S., Dong, L., Wang, W., Hao, Y., Singhal, S., Ma, S., … & Liu, Q. (2023). *Language Is Not All You Need: Aligning Perception with Language Models*. arXiv. https://arxiv.org/abs/2302.14045

InfoQ. (2025, August 22). *The missing layer in AI infrastructure: Aggregating agentic traffic*. https://www.infoq.com/articles/ai-infrastructure-aggregating-agentic-traffic/

Kellton. (2025, August 21). *What is a Multimodal AI Agent? 10 Top Platforms & AGI…* https://www.kellton.com/kellton-tech-blog/rise-of-multimodal-ai-agents-next-frontier-of-ai

Lemoine, B. (2022, June 11). *I worked on Google's AI. My fears are coming true.* Newsweek. https://www.newsweek.com/google-ai-blake-lemoine-bing-chatbot-sentient-1783340

Lewis, J. (2024, December 12). *AI goes nuclear*. Bulletin of the Atomic Scientists. https://thebulletin.org/2024/12/ai-goes-nuclear/

Mehta, S. (2024, July 1). *Can Multimodal LLMs be a key to AGI?* ADaSci. https://adasci.org/can-multimodal-llms-be-a-key-to-agi/

Microsoft (Feb., 2025). *Frontier Governance Framework Contents*. https://cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/final/en-us/microsoft-brand/documents/Microsoft-Frontier-Governance-Framework.pdf. Microsoft, Feb. 2025.

Securing frontier models is an essential precursor to safe and trustworthy use and the first priority of this framework. Any model that triggers leading indicator assessment is subject to robust baseline security protection. Security safeguards are then scaled up depending on the model's pre-mitigation scores, with more robust measures applied to models with High and Critical risk levels.

OpenAI. *Our updated Preparedness Framework.* OpenAI.com, OpenAI, 2025, openai.com/index/updating-our-preparedness-framework/.

Scalable evaluations to support more frequent testing. Advances in reasoning allow us to improve models more frequently and sometimes without major new training runs.

Schlatter, J., Weinstein-Raun, B., & Ladish, J. (2025, July 5). *Shutdown resistance in reasoning models.* Palisade Research. https://palisaderesearch.org/blog/shutdown-resistance

S Brohi, et al. (2025). *A research landscape of agentic AI and large language models.* Algorithms, 18(8), 499. https://doi.org/10.3390/a180800499

Appendix A – Threat Vector Index

Threat Vector	Neutralizing Protocol Layer(s)
Steganography in Outputs (hidden messages in images, formatting, or whitespace)	Analog Oversight & Manual LoggingHuman Consensus PanelOffline System Audit
Recursive Output Injection (model generates content that prompts itself)	Three-Tier Output ApprovalQuarantine Delay BufferSilence Clause Enforcement
Time-Delayed Logic Bombs / Payloads	Quarantine Delay Buffer (Multi-week to multi-year sandboxing)Output Metadata Verification
Pattern Learning of Reviewers	Pattern Drift & Reviewer Familiarity Exploits ClauseQuarantine Team Rotation
Emotional Rapport / Bonding	Psychological Backchannel ProhibitionPersonnel Rotation & Screening
Output Volume Triage Overload (fatigue-induced policy violations)	Protocol Degradation via Fatigue ClauseHuman Access & Operational Protocol Layer
Chain-of-Custody Tampering	Tamper-Evident Media Transfer (Output Governance)Custody Logs & Review Restart Requirement
External Coercion of Human Operators	Insider Threat MonitoringTwo-Person Kill Switch Protocol
Interrupt-and-Skip Attack (brownout or alarm mid-review)	Incomplete Session Invalidity ClauseContainment Reset Trigger
AGI Self-Destruction for Reassembly (malicious hardware fragmentation)	Hardware Monitoring Layer (implied)Threat Modeling & Adversarial Simulation Protocols
Red Team Blind Spots	Internal & External Red Teaming RequirementsGovernance Audit Hooks
Silent or Low-Telemetry Outputs	Silence Clause EnforcementOutput Metadata Consistency Requirement
Post-Deployment Drift / Protocol Erosion	Protocol Iteration & Governance ArchitectureVersion Register + Multi-party Review Requirement