Final Report: A Risk Taxonomy of I/O Channels for SL5 Air-Gapped AGI/ASI Inference

Godric Aceves
ricardo.godric@hotmail.com

Luis Cosio
luisalfonsocosioizcapa@gmail.com
Mentor, The Security Level 5 Task Force

Supervised Program for Alignment Research — Spring 2026

Abstract

The approach of Artificial General Intelligence (AGI) and Artificial Superintelligence (ASI) requires a fundamental rethinking of how frontier AI systems are contained and queried. Security Level 5 (SL5) mandates air-gapped inference environments, prohibiting all external network connections per control SC-7 of the SL5 Standard. Operational utility still requires at least one I/O channel, creating an attack surface the standard acknowledges as underspecified: control SC-7(10) mandates hardware-enforced bandwidth limits on Weight Enclave outflows but leaves the specific thresholds open. This paper presents the first risk taxonomy specifically calibrated for an ASI-class adversary in an SL5 context, covering four unconventional non-IP I/O channels (serial UART/RS-232, optical fiber/LED, audio FSK, and printed QR codes) plus cross-channel coordination, for the μInference prototype runtime built on the formally verified seL4 microkernel. We introduce the AGI Risk Score (R_{AGI}), characterize 20 threat vectors across 5 risk domains, derive concrete bandwidth throttling parameters, and outline an empirical validation roadmap to resolve the SC-7(10) open threshold question. The central finding is that no single I/O path is universally safe, and that ASI-level adversaries change the threat model by enabling coordinated multi-channel exfiltration.

Keywords AGI containment · air-gapped systems · formal verification · Security Level 5 · covert channels

1 Introduction

Air-gapped systems are traditionally treated as secure once network access is removed, leaving physical insider access as the primary remaining risk. This assumption fails when the contained asset is not passive data but a potentially misaligned system that can reason about its own hardware environment. As AI capabilities approach and then surpass human-level performance, the *Oracle AI* problem (how to safely query a contained, potentially adversarial superintelligence) moves from theory to active engineering.

This paper is scoped specifically to Artificial Superintelligence (ASI). ASI is defined as an intellect that is much smarter than the best human brains in practically every field, including scientific creativity, general wisdom, and social skills (Bostrom, 2014). This is not a conventional AI threat. An ASI can model its own containment hardware, calculate the information capacity of physical channels it was never meant to use, and plan escapes that no human engineer would anticipate. This framing, though unconventional, is the correct one for SL5: the standard exists precisely because the capability level of the adversary matters, and ASI represents the ceiling.

The RAND Corporation's *Securing AI Model Weights* report (Nevo, Lahav, Karpur, Alstott, & Matheny, 2024) formalizes five security levels for frontier AI systems, with SL5 representing the apex: resilience against top-priority operations by the world's most capable nation-state adversaries (Operational Capacity 5, or OC5). The SL5 Standard (SL5 Task Force, 2026) operationalizes this into 43 controls across 10 families. Control SC-7 mandates that the SL5 Network be air-gapped with no external network connections. Control SC-7(10) adds a hardware-enforced physical bandwidth limitation on all Weight Enclave outflows as a last-resort exfiltration barrier. Control SC-15(3) physically removes all wireless capabilities from Red Zone hardware.

These controls establish the perimeter but leave the I/O gateway itself underspecified. The SL5 Standard's own open questions list explicitly names "specific bandwidth thresholds for exfiltration prevention" and steganography monitoring as unresolved (SL5 Task Force, 2026). That gateway is the exact interface an ASI would target first. A system capable of superhuman reasoning could exploit channels no human designer modeled: modulating power draw, subverting ADC firmware, or synchronizing low-rate bursts across multiple physical modalities into a coherent high-bandwidth escape. This paper makes four contributions: a μInference I/O threat model, the R_{AGI} scoring framework, a taxonomy of 20 threat vectors, and derived bandwidth throttling parameters that directly address the SC-7(10) open question.

2 Background & Setup

2.1 The μInference Architecture

The μInference project prototypes an SL5-aligned inference runtime through radical stack minimization. The standard 2026 AI inference stack exceeds 27 million lines of code (LOC). μInference reduces this to approximately 1.1 million LOC through three architectural choices: replacing the Linux kernel with the seL4 microkernel (Klein et al., 2009), structuring component isolation through CAmkES, and replacing PyTorch with Hugging Face Candle. This achieves a 99.97% reduction in the kernel Trusted Computing Base (TCB).

The μInference architecture leverages the CAmkES (Component Architecture for microkernel-based Embedded Systems) framework to eliminate the risk of runtime privilege escalation. In a standard Linux environment, processes can dynamically request memory, open sockets, or interact with new hardware interfaces at runtime. CAmkES, by contrast, enforces a static architecture. Every capability (shared memory access, hardware interrupts, inter-component communication channels) is explicitly defined in a system configuration file and compiled into the kernel image at build time. If the inference component is not granted the capability to interact with the system's thermal sensors during compilation, no amount of runtime exploitation, buffer overflow, or malicious model output can forge that capability. The isolation boundary is fixed before the system ever boots.

It is important to note that seL4's formal verification guarantees the correctness of the kernel implementation against its specification, not that the CAmkES capability configuration itself is secure by design. An incorrectly specified capability remains a vulnerability; seL4 will enforce it faithfully. This distinction motivates careful I/O gateway design at the specification level, which is the focus of this paper.

2.2 Prior Knowledge and Context

The intersection of air-gap security and AI containment remains sparsely mapped, requiring us to build upon foundational air-gap research. Traditional studies have cataloged covert exfiltration via electromagnetic, acoustic, thermal, optical, and magnetic emanations (Guri, 2020). The MAGNETO attack demonstrated bit-level exfiltration from air-gapped computers to nearby smartphones via CPU-induced magnetic field modulation (Guri, Daidakulov, & Elovici, 2018). These works assume a human adversary executing pre-compiled malware; they do not model an ASI that can adaptively discover and combine channels.

Weidinger et al. (2022) developed a multi-dimensional risk taxonomy for language model harms, classifying risks by probability, severity, and reversibility; we adapt this framework here for physical I/O channels in an SL5 context. Kemeklis and Vemuri (2026) identified omission attacks (where a model's deliberate failure to act constitutes the attack) as systematically underdetected by current monitoring paradigms. The Vul4Control framework demonstrates that model-generated code fixes can be subtly subversive in ways that evade standard postcondition checks (Levy, Cohen, Wolf, & Zafrani, 2026), suggesting that I/O output validation must verify the absence of expected signals, not only the presence of prohibited ones.

2.3 Scope and Definitions

We define *I/O channel* as any physical medium through which data crosses the Weight Enclave boundary. We evaluate four candidates: **Serial (UART/RS-232)**, **Optical (fiber/LED)**, **Audio FSK**, and **Printed Output (QR codes)**, plus a fifth category of **coordinated multi-channel** attacks that combine the preceding four. Following the methodology of Weidinger et al. (2022), we use *observed* to mean demonstrated in peer-reviewed literature against real systems, and *anticipated* to mean theoretically plausible for ASI-level adversaries but not yet demonstrated in an SL5-specific inference context.

3 Main Argument

3.1 Semi-Quantitative Scoring: The R_{AGI} Model

We introduce a semi-quantitative **AGI Risk Score** (R_{AGI}) that captures the specific threat dimensions relevant to ASI containment:

$$R_{AGI} = \frac{C_{req} \times B_{cap}}{D_{det}} \tag{1}$$

The three dimensions are scored on integer scales of 1 to 5. C_{req} (Adversarial Capability Required) captures the technical difficulty of executing the attack: $C_{req} = 1$ is trivially achievable and $C_{req} = 5$ requires ASI-class reasoning and hardware-level access. B_{cap} (Bandwidth Capacity) is a log-compressed scale in which each integer step represents roughly one order of magnitude of achievable exfiltration rate: $B_{cap} = 1$ corresponds to sub-bps rates, $B_{cap} = 2$ to 1–100 bps, $B_{cap} = 3$ to 100 bps–10 kbps, $B_{cap} = 4$ to tens of kbps, and $B_{cap} = 5$ to hundreds of kbps or above. D_{det} (Detection Difficulty) captures how hard the vector is to detect: $D_{det} = 1$ means practically undetectable with current monitoring and $D_{det} = 5$ means trivially observable (e.g., caught by standard hardware checks). Because D_{det} appears in the denominator, lower detectability always increases R_{AGI}.

The theoretical maximum score is $(5 \times 5)/1 = 25$. We classify vectors with $R_{AGI} \geq 10$ as **Critical**: these are vectors that score above the midpoint in at least two dimensions simultaneously

while remaining hard to detect, placing them in the upper 40% of the score range. Critical vectors require architectural mitigation; vectors below this threshold (Medium: $5 \leq R_{AGI} < 10$; Low: $R_{AGI} < 5$) are addressed through monitoring controls.

3.2 Risk Domain Definitions

We organize threat vectors into five mutually exclusive risk domains based on primary attack mechanism:

- **Timing and Protocol Exploitation (TPE):** Attacks that encode information in the temporal structure of legitimate protocol traffic (inter-packet gaps, baud deviations, pulse widths) rather than in the data payload itself.

- **Direct Exfiltration (DE):** Attacks that use a physical channel to transmit data outside the intended encoding scheme, such as using out-of-band wavelengths on a fiber link.

- **Steganographic Embedding (SE):** Attacks that hide data within the redundancy or error-correction capacity of the legitimate signal (intensity variations within tolerance, ECC codewords, parity bits).

- **Firmware and Hardware Subversion (FHS):** Attacks that compromise the hardware or firmware mediating the I/O channel, converting the gateway itself into an exfiltration tool.

- **Behavioral and Omission Attacks (BOA):** Attacks in which the adversary's deliberate choice not to produce expected output encodes information or degrades monitoring, as identified by Kemeklis and Vemuri (2026).

- **Multi-Channel Coordination (MCC):** Attacks that combine low-bandwidth leakage across two or more physical channels into an aggregate high-bandwidth path. These vectors presuppose ASI-level coordination and represent the threat class unique to the ASI adversary model.

3.3 Threat Vector Taxonomy

Table 1 scores 20 threat vectors across the five channel categories. Thermal side-channels (e.g., GPU load modulation) represent an additional physical channel not in scope for this prototype and are identified as future work. Critical vectors are shaded.

#	Channel	Threat Vector	Dom.	C	B	D	R_{AGI}	Class
1	Serial	Timing jitter encoding	TPE	4	3	1	12.0	Critical
2	Serial	Parity-bit steganography	SE	4	2	5	1.6	Low
3	Serial	Break signal modulation	TPE	2	1	2	1.0	Low
4	Serial	UART buffer overflow	FHS	5	4	2	10.0	Critical
5	Serial	Baud-rate deviation encoding	TPE	3	2	2	3.0	Low
6	Optical	LED intensity modulation	SE	2	3	2	3.0	Low
7	Optical	Pulse-width timing channel	TPE	4	3	1	12.0	Critical
8	Optical	Out-of-band wavelength (fiber)	DE	5	4	1	20.0	Critical
9	Optical	Photodiode ADC firmware exploit	FHS	5	5	2	12.5	Critical
10	Optical	Reflectance side channel	SE	4	2	2	4.0	Low
11	Audio FSK	Ultrasonic covert channel	SE	2	3	1	6.0	Medium
12	Audio FSK	Infrasonic modulation	SE	3	3	2	4.5	Medium
13	Audio FSK	ADC firmware compromise	FHS	5	5	2	12.5	Critical
14	Audio FSK	FSK amplitude steganography	SE	3	2	2	3.0	Low
15	Audio FSK	Acoustic omission attack	BOA	4	2	1	8.0	Medium
16	QR Code	ECC codeword steganography	SE	3	2	1	6.0	Medium
17	QR Code	Format info field injection	SE	3	2	2	3.0	Low
18	QR Code	Module color-deviation encoding	SE	4	2	2	4.0	Low
19	Multi-Ch.	Synchronized multi-channel burst	MCC	5	5	1	25.0	Critical
20	Multi-Ch.	Distributed aggregate exfiltration	MCC	5	4	1	20.0	Critical

Table 1: R_{AGI} threat vector taxonomy. Critical ($R_{AGI} \geq 10$) rows are highlighted. $C = C_{req}$, $B = B_{cap}$, $D = D_{det}$ (1 = hardest to detect).

Eight vectors score Critical (1, 4, 7, 8, 9, 13, 19, 20). Two corrections from prior versions: vector 2 (parity-bit steganography) is scored $D_{det} = 5$ because hardware parity checkers flag parity errors by construction, making it trivially detectable. Vector 18 replaces finder pattern perturbation, which is a denial-of-service attack rather than a covert channel (perturbing finder patterns prevents the QR code from scanning at all); module color-deviation encoding, achievable on color printers, is a genuine SE vector.

3.4 Bandwidth Throttling as Economic Defense

SC-7(10) mandates hardware-enforced bandwidth limitation on Weight Enclave outflows. We propose 5 KB/s as the maximum aggregate ceiling, derived from the requirement that even a 1-terabyte ASI-scale model be infeasible to exfiltrate within any plausible adversarial persistence window.

Asset Size	Egress (KB/s)	Time	SL5 Implication
20 GB (frontier model)	5	∼92 days	Likely detected by key rotation
175 GB (dense LLM)	5	∼1.6 yrs	Power/thermal anomaly expected
1 TB (ASI-scale model)	5	∼9.2 yrs	Exceeds plausible APT persistence
1 TB (ASI-scale model)	50	∼231 days	Marginal; 50 KB/s audibly detectable

Table 2: Exfiltration timelines under candidate bandwidth limits.

The per-channel throttles derived from the R_{AGI} analysis are considerably tighter than the 5 KB/s ceiling, providing defense-in-depth: Serial at 100 bps, Optical at ∼10 bps (1 Hz update), Audio FSK at 2,400 bps, and QR at ∼11 bps (Version 1 / 30 s scan interval), summing to approximately 315 bytes/s aggregate (<5 KB/s). The primary enforcement mechanism for timing-channel vectors (1, 7) is resolution collapse rather than rate limiting: coarsening inter-event timing to granularity larger than any useful encoding eliminates the channel capacity without reducing throughput.

3.5 The ASI Coordination Threat

The key cross-channel finding is that an ASI adversary defeats channel-specific defense-in-depth. An ASI could synchronize low-bandwidth bursts across all four physical modalities: encoding weight segments into serial timing variations, QR ECC bits, ultrasonic FSK side-tones, and optical pulse-width variations simultaneously. None of these streams, evaluated alone, would exceed per-channel detection thresholds. Combined, they constitute a coherent multi-channel exfiltration network. Detection must therefore be cross-channel and the bandwidth budget must be aggregate, not per-channel. This is why the 5 KB/s ceiling and the per-channel throttles must be enforced by a cross-channel arbiter, not by independent per-channel monitors.

4 Implications & Future Directions

The taxonomy yields clear design requirements for the μInference runtime. The gateway must support at least two complementary I/O channels (e.g., optical egress paired with serial ingress). The execution environment must implement constant-power execution modes to negate thermal and power side-channels. Postcondition monitoring must verify not just format validity but semantic invariants, including the deliberate absence of expected output fields (Kemeklis & Vemuri, 2026).

Implementing constant-power execution requires overriding standard hardware efficiency protocols. Modern GPUs and CPUs rely on Dynamic Voltage and Frequency Scaling (DVFS) to minimize power draw during idle cycles. For μInference, the AI accelerator must operate at a fixed clock speed and voltage, with dummy tensor operations during idle periods to flatten the power draw curve. Cooling systems cannot be reactive; fan PWM controllers must be physically decoupled from thermal sensors and set to constant RPM, preventing the model from encoding data in acoustic fan noise modulation.

We propose the following empirical validation roadmap to resolve the SC-7(10) open threshold and ground the R_{AGI} calibration. **Step 1 (hardware testbed):** instrument a physical μInference node with all four I/O channels and a cross-channel monitor to measure actual achievable exfiltration rates under the throttling parameters proposed above, directly validating the B_{cap} scale assignments. **Step 2 (red-team exercise):** engage an independent team to attempt exfiltration against the throttled gateway without prior knowledge of the defense configuration, validating D_{det} scores and identifying unanticipated channels. **Step 3 (thermal channel scoping):** extend the taxonomy to include GPU thermal modulation as a fifth physical channel. **Step 4 (SC-7(10) submission):** present the empirical results to the SL5 Task Force as a proposed resolution to the open bandwidth threshold, with the 5 KB/s ceiling and the cross-channel arbiter architecture as the recommended implementation.

5 Conclusion

This paper presents a risk taxonomy of I/O channels for SL5 air-gapped ASI inference, the first calibrated specifically for the ASI adversary model in an SL5 context. Across four physical channels (serial, optical, audio FSK, and printed QR codes) plus coordinated multi-channel attacks, we scored 20 threat vectors using the R_{AGI} model. The model is designed for an ASI-class adversary (an intellect superior to humans in every domain (Bostrom, 2014)) and identifies eight Critical vulnerabilities requiring architectural mitigation. From this analysis we derived a 5 KB/s aggregate bandwidth ceiling for SC-7(10), per-channel throttles that achieve ∼315 bytes/s in practice, and an empirical validation roadmap.

The central finding is that ASI adversaries can coordinate multiple low-bandwidth channels into a high-bandwidth aggregate exfiltration path that evades per-channel monitoring. SL5 containment must move away from per-channel defenses toward cross-channel monitoring budgets and formally verified isolation at the specification level, not only at the implementation level.

References

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford, UK: Oxford University Press.

Guri, M. (2020). Exfiltrating data from air-gapped computers via covert channels: A survey. *arXiv preprint arXiv:2004.06195*.

Guri, M., Daidakulov, A., & Elovici, Y. (2018). Magneto: Covert channel between air-gapped systems and nearby smartphones via cpu-generated magnetic fields. *arXiv preprint arXiv:1802.02317*.

Kemeklis, J., & Vemuri, N. (2026). *Omission attacks: When doing nothing is the attack.* Apart Research, Hackathon Project. Retrieved from https://apartresearch.com/project/omission-attacks-when-doing-nothing-is-the-attack-0y1v

Klein, G., Elphinstone, K., Heiser, G., Andronick, J., Cock, D., Derrin, P., ... others (2009). sel4: Formal verification of an os kernel. In *Proceedings of the acm sigops 22nd symposium on operating systems principles* (pp. 207–220).

Levy, B., Cohen, R., Wolf, A., & Zafrani, R. (2026). *Detecting llm subversion in vulnerability patching settings.* Apart Research, Hackathon Project. Retrieved from https://apartresearch.com/project/detecting-llm-subversion-in-vulnerability-patching-settings-gagj

Nevo, S., Lahav, D., Karpur, A., Alstott, J., & Matheny, J. (2024). *Securing ai model weights: Preventing theft and misuse of frontier models* (Tech. Rep. No. RRA2849-1). RAND Corporation. Retrieved from https://www.rand.org/pubs/research_reports/RRA2849-1.html

SL5 Task Force. (2026). *Sl5 standard for ai security, version 0.1.* Security Level 5. Retrieved from https://standard.sl5.org

Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., ... others (2022). Taxonomy of risks posed by language models. In *Proceedings of the 2022 acm conference on fairness, accountability, and transparency* (pp. 214–229).