**A-SWARM Autonomy Guardrails & Safety — Rules of Engagement (One-Pager)**

**Version 1.0 — August 19, 2025**

**Purpose**

Establish non-negotiable safety, security, and governance boundaries for A-SWARM deployments in AI datacenters and OT/ICS environments. This document defines what A-SWARM **will** and **will not** do—technically, procedurally, and legally.

**Scope of Operation**

* **Assets covered:** Compute hosts (Linux/Windows), network segments, hypervisors, storage, and PLC-adjacent gateways.
* **Explicitly out of scope:** Safety-Instrumented Systems (SIS) and safety PLC logic (SIL loops), breaker/protection relay settings, turbine/governor set points, and any change that bypasses physical interlocks.

**Autonomy Rings (Actuation Boundaries)**

1. **Ring 0 — Observe (Default):**  
   Read-only sensing, anomaly detection, advisory alerts, recommended playbooks. **No actuation.**
2. **Ring 1 — Micro-Actuation (Pre-authorized, Reversible):**  
   Limited, bounded actions on IT/edge assets only, e.g.:
   * Rate-limit/shape traffic on a host or ToR port
   * Quarantine a VM/VLAN / disable a single interface
   * Kill/contain a non-safety process / revoke a token / rotate a key
   * Block an IOC at a local firewall  
     **All Ring-1 actions:**
   * Must be pre-approved in writing per site/runbook
   * Are time-boxed and auto-revert unless explicitly extended
   * Produce a signed **Action Certificate** (see “Audit & Forensics”)
3. **Ring 2 — Major Actuation (Quorum-Gated, Human-in-Loop):**  
   Segment isolation, rack/pod isolation, BGP/route withdrawals, OT cell isolation.  
   **Never autonomous.** Requires multi-party approval (Ops + Sec + Duty Manager), physical/dual-control confirmation, and a maintenance window unless in declared emergency.
4. **Ring S — Safety/Protection Systems:**  
   **Never touched by A-SWARM.** Read-only telemetry via certified taps only.

**Twin-First Policy (Adversarial R&D)**

* All attack synthesis, red teaming, and policy learning run **exclusively** inside a **per-site digital twin** seeded with production telemetry.
* No “hacking back,” no live offensive tests, no payloads or exploits on production systems.
* Twin drift is monitored; no policy moves to prod without a successful twin rehearsal.

**Safety Interlocks & Kill-Switches**

* **Site-Local Kill:** Physical/EoL button or out-of-band signed command halts all A-SWARM actuation within that site; Ring-0 monitor remains available for visibility.
* **Global Suspend:** Organization-wide freeze of any actuation class (e.g., disable all quarantine actions) via HSM-signed instruction.
* **Floor/Ceiling Limits:** Per-asset and per-site hard caps on rate limits, quarantine scope, and duration; enforced locally even during controller loss.

**Determinism, Verification & Change Control**

* **Deterministic Agents:** Bounded CPU (<1%), RAM (<50 MB), and egress (<5 kbps avg); finite-state policy engines with reproducible outcomes.
* **Policy Compiler:** Human-readable YAML → **policy automata** with machine-checkable safety/liveness constraints.
* **Canary & Revert:** Blue/green rollout, canary scope <5% of assets; auto-rollback on violation triggers.
* **Time-Sync:** PTP/NTP normalization for consistent replay and decision logs.

**Security & Trust**

* **Supply Chain & Runtime:** Signed/attested builds (Secure Boot, TPM/TEE/SEV-SNP/TDX where available); least privilege; no shell access to SIS/Safety PLCs.
* **Crypto:** PQC for transport keys; per-site HSM for key custody; forward-secure log sealing.
* **Resilience:** Byzantine-tolerant “pheromone” gossip; rate-limited propagation; partition-safe defaults (fail-quiet at Ring-1).

**Audit & Forensics (Action Certificates)**

Every Ring-1/2 action emits an immutable **Action Certificate** containing:

* Time-sync stamps; asset IDs; telemetry digests; anomaly graph hash
* Policy version & proof hash; approvers (if Ring-2); scope & TTL
* Outcome/rollback; operator notes; cryptographic signatures  
  Certificates are exportable to SIEM/SOAR and form the chain-of-custody for regulators/insurers.

**Data Governance & Sharing**

* **Ownership:** Customer owns all telemetry and derived analytics.
* **Processing:** Minimum necessary data; privacy by design; role-based access; on-prem-first processing.
* **Cross-Site “Antibody” Sharing:** Opt-in, privacy-preserving exchange of de-identified IOCs/behavioral signatures with differential privacy budgets and cryptographic provenance.

**Compliance & Legal**

* **Controls Mapping:** Evidence packs aligned to IEC 62443, NIST 800-82, and (where applicable) NERC CIP; autogenerated control-by-control artifacts.
* **No Hackback:** A-SWARM does not perform offensive actions, counter-intrusions, or any activity that could be construed as “hacking back.”
* **Change Management:** ITIL-aligned; maintenance windows honored; plant manager override respected at all times.

**Deployment Interfaces & Integrations**

* **ICS/OT:** Read-only adapters for Modbus/TCP and OPC UA (v1); vendor-certified gateway sidecar for PLC adjacency—no code on safety PLCs.
* **IT:** SPAN/mirror ingestion, NetFlow/sFlow, syslog, EDR/SIEM connectors (Splunk/Elastic/CrowdStrike/Palo Alto/Microsoft).

**Operability SLOs (Pilot Targets)**

* **MTTD (coordinated anomaly):** < **200 ms**
* **MTTR to micro-containment (Ring-1):** < **5 s**
* **False-positive rate at Ring-1:** < **0.5%** (no false majors)
* **Auto-revert on no-harm principle:** default restore ≤ **60 s**

**Shared Responsibility (Summary)**

* **Customer:** Approves pre-auth Ring-1 catalog; designates quorum approvers; maintains time-sync and tap integrity; owns incident command.
* **A-SWARM:** Provides deterministic agents, policy compiler/proofs, twin build & rehearsal, audit artifacts, compliance mapping, and 24/7 support.

**Acknowledgement:** By enabling A-SWARM beyond Ring-0, the customer confirms these guardrails, interlocks, and procedures are in force and that any Ring-2 actuation requires documented human approval per this One-Pager.