MASO

Multi-Agent Security Operations

Practitioner Training Course

Enterprise AI Security Framework

Jonathan Gill · airuntimesecurity.co.za

Course Structure

Five modules. Building blocks. Each one depends on understanding the one before it.

01

Why Multi-Agent is Different

The problem single-agent controls don't solve

02

Risk Tiering for Agentic Systems

Classify before you control

03

Infrastructure Foundations

Message bus, NHI, guardrails — what to build first

04

PACE Resilience

How things fail. How you recover. How you degrade gracefully.

05

The MASO Architecture

Six control domains. Three tiers. Putting it all together.

MODULE 01

Why Multi-Agent is Different

The failure modes that single-agent controls will never catch.

Four Problems You Don't Have With Single Agents

Poisoned Handoffs

A compromised document processed by one agent becomes instructions for the next. Indirect prompt injection propagates through the chain.

Transitive Authority

If Agent A can delegate to Agent B, and Agent B has write access, then Agent A effectively has write access. Permissions leak through delegation.

Hallucination Amplification

Agent A hallucinates a claim. Agent B cites it. By Agent C, it's elaborated and presented with high confidence. Errors compound, not cancel.

Failures That Look Like Success

Three agents agreeing doesn't mean three independent opinions — not when they share the same model, training data, and retrieval corpus.

The Core Insight

Single-agent controls assume the agent is the boundary.
In multi-agent systems, the boundary is the communication between agents — and that's where attacks live.

This is why MASO exists. It secures the spaces between agents, not just the agents themselves.

MODULE 02

Risk Tiering for Agentic Systems

Classify the agent system before selecting controls. Not everything needs Tier 3.

Three Implementation Tiers

Start at Tier 1. Earn your way up. The tier determines how much autonomy agents get.

TIER 1

SUPERVISED

Autonomy

Human approves all writes

Key Controls

Guardrails active
Message bus in logging mode
Judge optional
Manual review mandatory

Use When

Pilots, regulated data, no operational baselines, first deployment

TIER 2

MANAGED

Autonomy

Auto-approve low-risk
Escalate high-risk

Key Controls

NHI per agent
Signed message bus
LLM-as-Judge active
Anomaly scoring
PACE A/C configured

Use When

Established baselines, proven controls, operational evidence

TIER 3

AUTONOMOUS

Autonomy

Minimal human intervention

Key Controls

Self-healing PACE
Adversarial testing
Independent observability agent
Kill switch tested

Use When

Mature ops, proven resilience, full PACE tested under stress

How to Classify an Agent System

Ask these five questions. If you answer "yes" to any, start at Tier 1.

1	Does the agent have write access to customer-facing systems or regulated data?
2	Is this the organisation's first multi-agent production deployment?
3	Could an uncorrected agent error cause financial loss, regulatory exposure, or reputational damage?
4	Does the agent chain cross trust boundaries (e.g., internal data → external API)?
5	Is the AI security maturity below CMMI Level 3 for AI-specific controls?

Tier 1 is not a permanent state. It builds the evidence base that justifies moving to Tier 2.

MODULE 03

Infrastructure Foundations

Three things to build before you deploy a single multi-agent workflow.

The Three Infrastructure Pillars

Secure Message Bus

All agent-to-agent communication routes through this bus.
No direct agent-to-agent communication permitted.
At Tier 1: logging mode (capture everything).
At Tier 2+: cryptographically signed messages.
DLP inspection on the bus catches data leakage between agents.

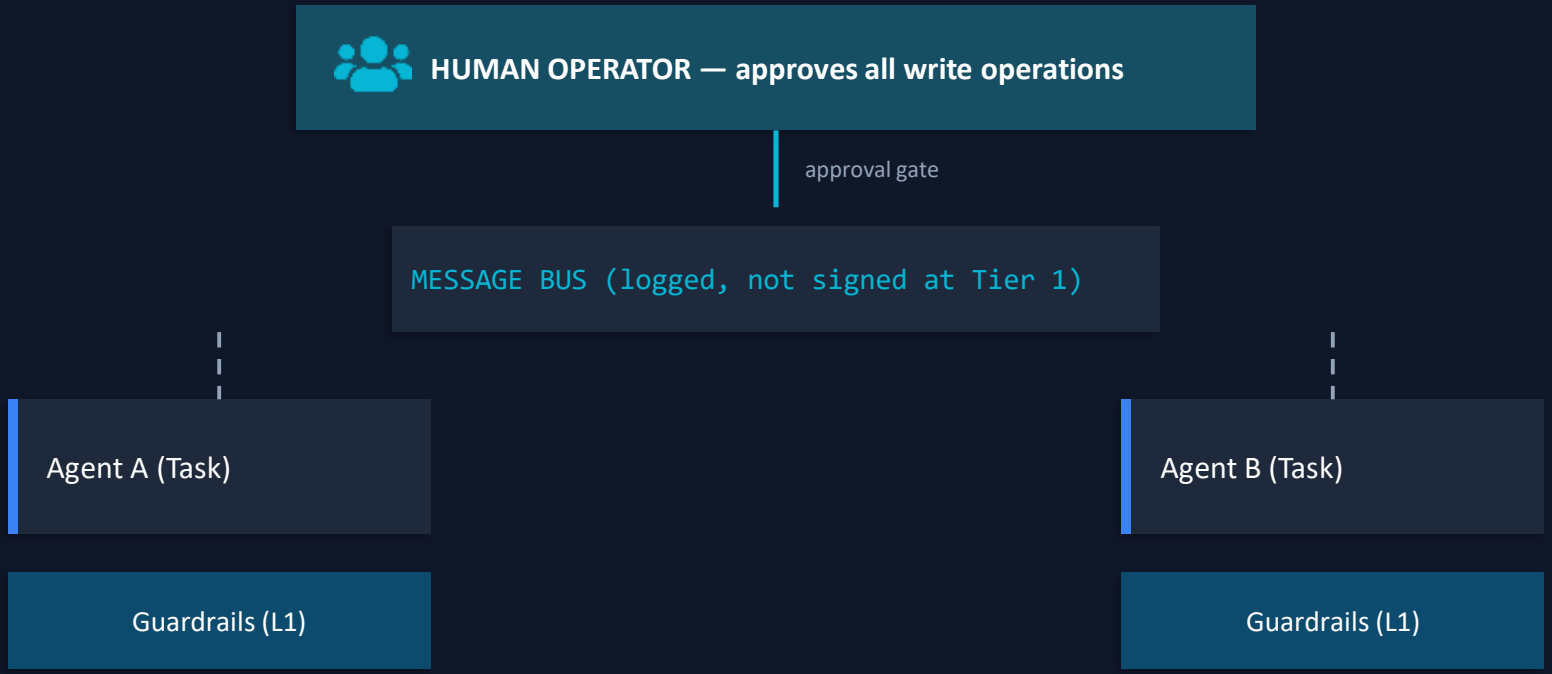
Non-Human Identity (NHI)

Every agent instance gets a unique identity in your IdP.
Short-lived credentials. No shared API keys.
Every agent has a human sponsor accountable for its actions.
Least privilege by default. Start with zero permissions.
Quarterly access reviews — same cadence as human accounts.

Guardrails (Layer 1)

Deterministic controls on every agent. Non-negotiable.
Input validation, output sanitisation, tool permission scoping.
Rate limiting per agent and per chain.
Operates at ~10ms. Catches known-bad patterns.
This is your floor. Everything else builds on this.

Tier 1 Architecture — What It Looks Like



Layer 2 (LLM-as-Judge): OPTIONAL at Tier 1
Layer 3 (Human Oversight): MANDATORY for all writes

MODULE 04

PACE Resilience

Primary. Alternate. Contingency. Emergency.
How to plan for failures before they happen.

PACE for Multi-Agent Systems

Military-grade resilience planning. Every layer trades capability for survivability.

P — **PRIMARY** — Normal ops

All agents operational. Full autonomy within tier permissions. Message bus integrity verified. Judge evaluates cross-agent outputs.

A — **ALTERNATE** — Degraded

One agent anomalous. Isolate it. Activate backup (different provider if possible). Tighten tool permissions to read-only across chain. Human approves all writes.

C — **CONTINGENCY** — Minimal

Multiple agents degraded or correlated failure detected. Suspend multi-agent orchestration entirely. Human approves every action. Reduced capacity, high assurance.

E — **EMERGENCY** — Full stop

Confirmed compromise or cascading failure. Kill switch fires. All agent sessions terminated. Non-AI fallback path activated. Incident response engaged.

Why Multi-Agent PACE is Harder

Blast radius is wider

A compromised agent can inject instructions into the message bus affecting every downstream agent. Containment must isolate the agent and quarantine its recent outputs across the entire chain.

Transitions must be automated at Tier 2+

Multi-agent cascading failures move faster than human response times. The monitoring agent or orchestrator initiates P→A transitions without waiting for human approval. Humans are notified, not gated.

Recovery requires chain verification

Stepping back from C→A isn't just 'restart the failed component.' You must verify no poisoned data persists in other agents' memory, context, or RAG corpus.

The concurrent advantage

Unlike traditional PACE (one layer at a time), your guardrails and Judge run in parallel on every request. Human oversight is on standby. You're continuously verifying through multiple independent channels.

MODULE 05

The MASO Architecture

Six control domains. 93 controls. 99 tests.
The complete picture.

Six Control Domains

Prompt, Goal & Epistemic Integrity

Agent instructions, objectives, information quality across chains. Injection, goal hijack, groupthink, hallucination amplification, uncertainty stripping.

Identity & Access

Non-Human Identity per agent, zero-trust credentials, scoped permissions, no transitive authority.

Data Protection

Cross-agent data fencing, DLP on the message bus, RAG integrity, memory isolation between agents.

Execution Control

Sandboxed execution, blast radius caps, circuit breakers, LLM-as-Judge gate, interaction timeouts.

Observability

Decision chain audit, anomaly scoring, drift detection, independent observability agent with kill switch.

Supply Chain

AIBOM per agent, signed tool manifests, MCP server vetting, A2A trust chain validation.

Domain 0 — Prompt, Goal & Epistemic Integrity

Input Sanitisation

All channels, not just user-facing. Strip injection patterns from inter-agent messages.

System Prompt Isolation

Prevent cross-agent extraction. Each agent's instructions are opaque to other agents.

Goal Integrity Monitoring

Immutable task specs. Continuous verification that agent objectives haven't drifted or been hijacked.

Anti-Groupthink Controls

Detect correlated outputs from agents sharing model/data. Force independent evaluation.

Uncertainty Preservation

Track confidence qualifiers across handoffs. Flag when 'possibly' becomes 'confirmed' without basis.

Domain 1 — Identity & Access

Non-Human Identity (NHI)

Unique identity per agent in your IdP. Tagged as AI agent with risk tier and owning team.

Zero-Trust Credentials

Short-lived tokens (OAuth/OIDC).
No long-lived API keys.
Automatic rotation.

Scoped Permissions

Least privilege by default.
Start with zero. Grant only what's declared for purpose.

No Transitive Authority

Agent A delegating to Agent B must not inherit B's permissions.
Each agent authenticates independently.

Human Sponsor

Every agent identity has an accountable human. Quarterly reviews same as human accounts.

Domain 2 — Data Protection

Cross-Agent Data Fencing

Prevent uncontrolled data flow between agents at different classification levels.

Message Bus DLP

Output DLP scanning on inter-agent comms. Catches PII, secrets, and sensitive data in transit.

RAG Integrity Validation

Verify knowledge base hasn't been tampered with. Hash checks on retrieval corpus per query.

Memory Isolation

No shared memory between agents unless explicitly declared.
Memory poisoning detection active.

Domain 3 — Execution Control

Sandboxed Execution

Every tool invocation runs in isolation. Strict parameter allow-lists. Filesystem/network scoping.

Blast Radius Caps

Limit damage any single agent can do before circuit breakers engage. Rate + scope limits.

Circuit Breakers

Auto-trigger PACE escalation when error rates exceed thresholds. Fail-closed by default.

LLM-as-Judge Gate

Independent eval model approves actions before commit. Distinct from task agents.

Interaction Timeouts

Hard time limits on agent sessions and inter-agent exchanges. Prevents infinite loops.

Domain 4 — Observability

Decision Chain Audit

Immutable logs of full reasoning and action history. Every agent, every decision, every handoff.

Anomaly Scoring

Per-agent behavioural scoring. Feeds into PACE escalation logic. Baseline comparison continuous.

Drift Detection

Compare current agent behaviour against established baselines. Flag deviations before impact.

Independent Observer Agent

Separate agent on different model/ provider. Monitors the chain. Owns the kill switch.

Domain 5 — Supply Chain

AIBOM per Agent

AI Bill of Materials for each agent. Model version, training data provenance, dependency tree.

Signed Tool Manifests

Every tool an agent can invoke has a signed, versioned manifest. No unsigned tool execution.

MCP Server Vetting

Model Context Protocol servers audited and allow-listed before agents can connect.

A2A Trust Chain

Agent-to-Agent protocol validation. Verify identity and integrity of every agent in the chain.

What Makes MASO Different

EPISTEMIC SECURITY

Epistemic = relating to knowledge — how information is acquired, validated, and trusted.

Think of it as the broken telephone problem at machine speed. When agents pass information between themselves, facts get distorted, confidence inflates, and qualifiers disappear — not because anyone attacked the system, but because that's what happens in a chain.

Groupthink	Agents converging on the same wrong answer because they share model, training data, and retrieval corpus
Correlated Errors	Failures that look independent but aren't — same underlying bias, same data, same blind spots
Synthetic Corroboration	Agent B citing Agent A's hallucination as evidence. By Agent C, it's 'well-established'
Uncertainty Stripping	Each handoff strips qualifiers. 'Possibly X' becomes 'likely X' becomes 'X is confirmed'

The Three-Layer Defence Model

Every request passes through all active layers. They run concurrently, not sequentially.

L1 — GUARDRAILS — ~10ms

Deterministic. Non-negotiable. Machine-speed.

Input validation, output sanitisation, tool scoping, rate limiting.

Catches known-bad: blocked topics, PII, schema violations, injection patterns.

L2 — LLM-AS-JUDGE — ~500ms–5s

Independent evaluation model (distinct from task agents).

Assesses quality, safety, policy compliance before outputs commit.

In multi-agent: evaluates inter-agent comms for goal integrity and injection.

Catches unknown-bad: policy drift, hallucination, subtle manipulation.

L3 — HUMAN OVERSIGHT — Min–Hours

The governance backstop. Scales inversely with demonstrated trustworthiness.

Write ops, external API calls, irreversible actions escalate based on risk.

Catches edge cases: ambiguous outputs, regulatory judgment, novel scenarios.

MASO Coverage at a Glance

93	99	3	50
Controls	**Tests**	**Tiers**	**Risks Mapped**
across 6 domains	verification criteria	progressive autonomy	OWASP + emergent

Risk Coverage

OWASP LLM Top 10 (2025)	10/10
OWASP Agentic Top 10 (2026)	10/10
Emergent risks (no OWASP equivalent)	30

Full risk mapping available at airuntimesecurity.co.za/maso

Getting Started — Monday Morning Actions

Week 1

Inventory & Classify

Map every multi-agent workflow in production or planned.
Classify each using the five-question test.
Assign a human sponsor to every agent identity.

Week 2

Build the Floor

Deploy guardrails (Layer 1) on all agents — input, output, tool scoping.
Enable message bus logging (not signing yet).
Provision unique NHI per agent in your IdP.

Week 3

Plan for Failure

Define your PACE posture for each workflow.
Document the kill switch path. Test it.
Identify your non-AI fallback for each agent chain.

Week 4

Baseline & Iterate

Review first month of bus logs and guardrail data.
Establish behavioural baselines per agent.
Decide: stay Tier 1 or begin Tier 2 prep.

MASO

Multi-Agent Security Operations

---

airuntimesecurity.co.za

Open source · MIT License · Enterprise AI Security Framework

Jonathan Gill