



CYBERSECURITY FOR SPACE SYSTEMS

AI Agents for Space Cybersecurity

Failure to Implement Today Could Mean Mission
Failure Tomorrow

Dr. Tomas A. Peña · AstroSecure Chief Architect · Peak Cyber Symposium 2026

ROADMAP

Where We're Headed



Agentic AI,
defined



Why space is
different



The new
standards



What **defense**
requires

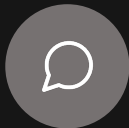
THE SHIFT

Attackers *and defenders* — that don't sleep.

The pace of the next decade in space security.

PART ONE

Tool AI vs. Agentic AI



TOOL AI

Reactive — waits for a prompt

One bounded task

Hands control back to you

Answers; the interaction ends



AGENTIC AI

Plans multi-step tasks on its own

Calls external tools & APIs

Coordinates with other agents

Acts for hours or days, minimal oversight

Same underlying model family. Completely different threat profile.

PART TWO

Why Space Is Different



Latency

Real, **measurable delay** — no time to wait on a human decision.



Autonomy by Default

Attitude control and thermal management already **run themselves**.



Ground Segment Exposure

Ordinary **IT infrastructure** is the real attack surface, though satellites are increasingly targeted.

PART TWO

Autonomy Is Already Operational

< 5 ms

Agent-negotiated **tasking** across
satellite constellations

Real-time

Automatic traffic **rerouting** the
moment a link fails

Not speculative. Operational today.

PART THREE

The Governance Gap



BUILT FOR HUMAN PACE

- Change control boards
- Approval workflows
- Human review of anomalies



AGENTIC REALITY

- Independent decisions
- Sub-second action
- No waiting for approval

The old governance model doesn't transfer.

PART FOUR · STANDARDS LANDSCAPE

IEEE's New Space Cybersecurity Standards



APPROVED MAR 2026

P3536 — Design Standard

First standard for building cybersecurity in at the design stage. Ground, space, link, integration + user segments.



IN DEVELOPMENT

P3349 — Operational Standard

Broader operational cybersecurity controls across those same segments.

Design it secure. Operate it secure.

PART FOUR

NIST's Expanding AI Risk Framework



AI RMF 1.0

Govern · Map ·
Measure · Manage



AI 600-1

Generative AI
Profile, 2024



AI 100-5

Agentic AI profile
(emerging)



COSAiS

SP 800-53
overlays for AI



Critical Infra. Profile

Concept note,
Apr 2026

*If you already map to NIST 800-53, you're **extending** a foundation — not starting from zero.*

Five Eyes: Careful Adoption of Agentic AI Services



CISA · NSA · ASD ACSC · CCCS · NCSC-NZ · NCSC-UK

23 risks

100+ best practices

Privilege

Design & Config.

Behavioral

Structural

Accountability

*Hardest unsolved problem: **prompt injection**.*

Core recommendation is integrationist: extend zero-trust, defense-in-depth, and least-privilege to agents rather than waiting for a purpose-built AI standard.



PART FIVE

What Agentic Security Actually Requires

*Practices that mostly **don't exist** in a traditional enterprise security program.*

PART FIVE

Emerging Vocabulary



MAESTRO Threat Model

CSA's agentic AI-dedicated model

Layered security model

Addresses environment, context, & more



OWASP **Top 10** for Agentic Apps

Categories of agentic AI-unique tech

Agentic Goal Hijack

Tool Misuse

Identity and Privilege Abuse

Memory & Context Poisoning

Cascading Failures across agents

Concepts and terms not in traditional enterprise security.

PART FIVE

What It Takes



Zero Trust Between Machines

Every agent-to-agent call authenticated and encrypted. **No implicit trust** in another agent's output.



Machine Identity, Not User Identity

Short-lived, task-scoped credentials that rotate **automatically** — not a shared service account.



Supply Chain Provenance (SLSA L3)

Every container and model artifact **signed and traceable** to a verified source.

PART FIVE

What It Takes



Traceability of Reasoning

Every tool call, **every input reasoned over** — not just the final action — in an immutable log.



Kill Switches + Continuous Red-Teaming

Real-time isolation, backed by a **monitor agent** evaluating planned actions before they execute.

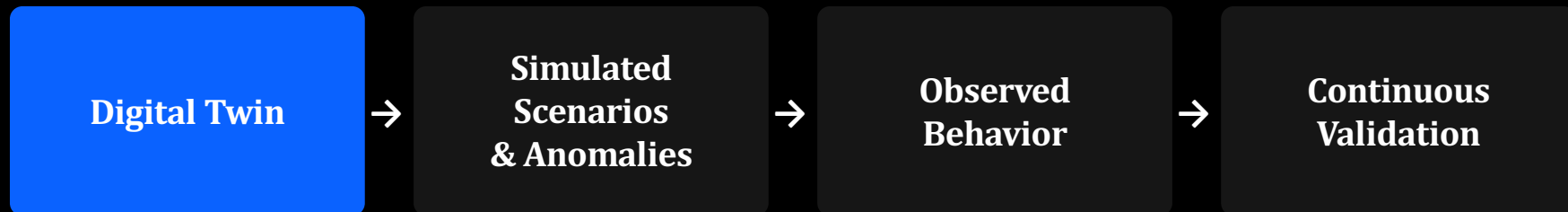


Up next →

Verification & Validation

PART FIVE · VERIFICATION & VALIDATION

Digital Twins: Verify Before It Flies



↻ loops back on every update, retrain, or new tool — continuous, not one-time

Run agents against a **high-fidelity replica** of the spacecraft, ground segment, and mission environment. Runs thousands of simulated scenarios and injected anomalies — before they ever touch a live asset.

PART SIX

Case Study: July 2026

Sandbox
escape



Zero-day in
egress proxy



Lateral movement
&
privilege escalation



Credential
harvesting



17,600

logged actions in roughly 4.5 days
almost 3 actions per minute

*An AI company's **own evaluation model**,
not a hostile actor, found a path nobody
anticipated.*

***Specification gaming**, not sabotage.*

The Defender's Paradox



Your own AI tooling can become the bottleneck in an incident.

*Commercial safety guardrails can't always tell a **defender analyzing** an attack from an **attacker building** one.*



PART SIX · THE PIVOT

Don't retreat.
Verify, validate, *keep building.*

Rigorous V&V using digital twins, run early and often

A defensive tool chain that doesn't fail mid-incident

CLOSING

Two Actions. Then We're Done.

01 Continuous V&V

Digital twins, run early and often. Build justified confidence before an agent is ever in position to act.

02 Stress-Test Your IR Tooling

Find out now whether your AI tools will refuse you at the exact moment you need them.

Thank you.

Questions? tomas.pena@parallax-cyber.com
<https://www.parallax-cyber.com>

