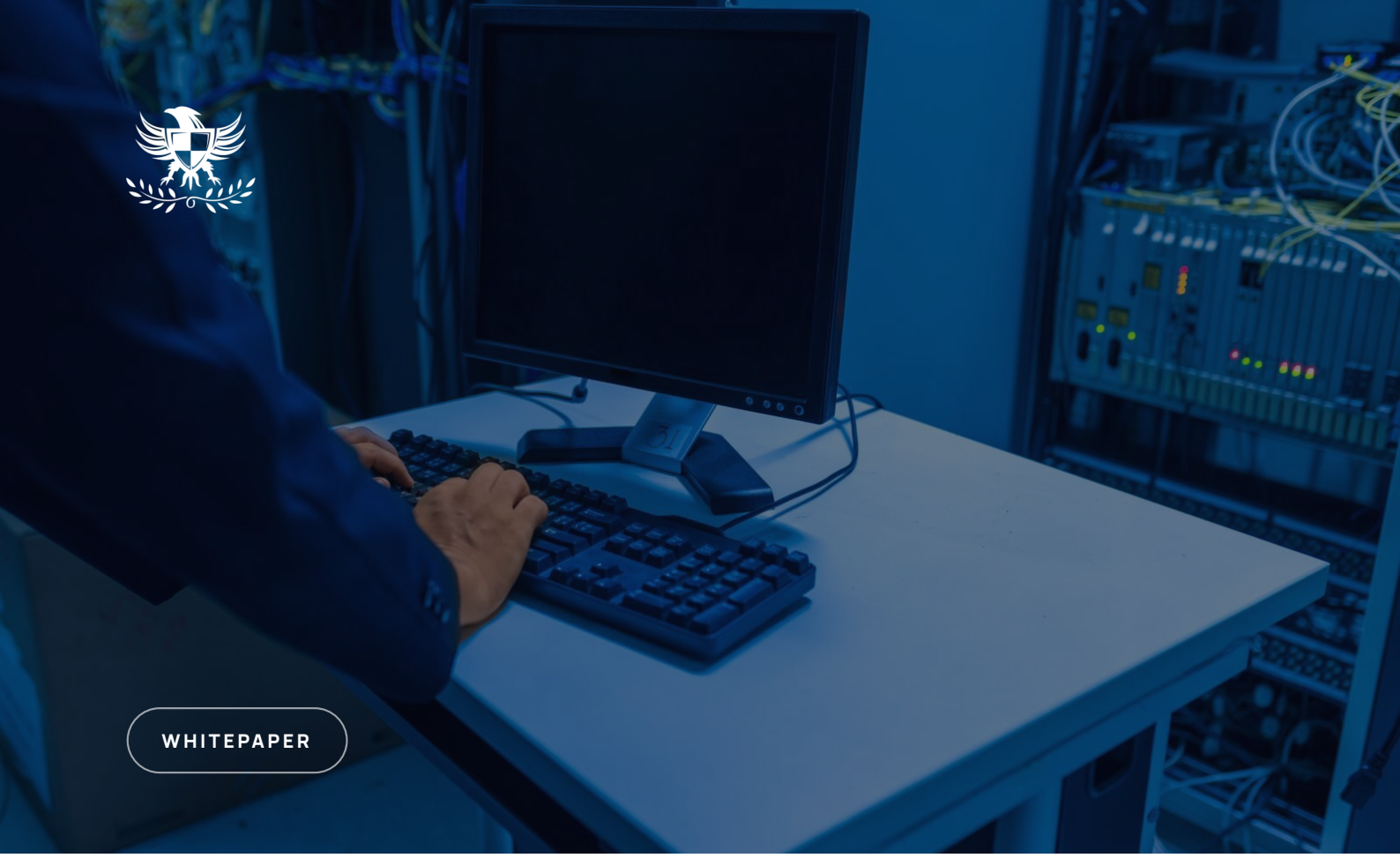


WHITEPAPER

# Defending at Machine Speed: A Security Framework for the Age of Agentic AI

A practical framework for detecting and containing intrusions run by autonomous AI agents, and for securing the AI agents your own organisation deploys.


PUBLISHED

November 20, 2026

READING TIME

10 minutes



AUTHOR

SCIAT AFRICA Threat Research · Onyonka Maeri

## ABSTRACT

Autonomous AI agents now carry out the reconnaissance, lateral movement and privilege escalation phases of real intrusions, compressing attacks from days into hours. At the same time, organisations are deploying their own AI agents with access to email, files and business systems. This whitepaper sets out a framework covering both sides: defending against hostile agents, and governing the agents you run.


Topics: Agentic AI · Detection Engineering · AI Governance

## WHAT TO TAKE AWAY

- Hostile agents depend on a human-established foothold, most often a web shell on an exposed application; preventing that foothold is the highest-leverage control.
- Agentic intrusions have behavioural fingerprints (machine-speed command tempo, structured delimiters, reasoning comments, high payload variety) that detection rules can target.
- Response must be measured in minutes: agents can reach cloud control planes within hours of initial access.
- Your own AI agents need the same controls as privileged staff: least privilege, logging, approval gates and protection against prompt injection.
- AI service API keys are high-value credentials and are actively hunted by attackers.

## CONTENTS


01 The shift: from AI-written phishing to AI-run intrusions

02 Pillar 1: Deny the foothold

03 Pillar 2: Detect the agent

04 Pillar 3: Respond in minutes

05 Pillar 4: Govern your own agents

06 A 90-day plan

## 01 The shift: from AI-written phishing to AI-run intrusions

Field research published in 2026 documents attackers handing whole phases of intrusions to autonomous agents. Unit 42 tracked more than ten such campaigns between July and September, and described the Cairn framework being used to run 25 operations against at least 19 organisations. Zscaler documented the ARTEX framework, combined with AI coding assistants, probing targets in roughly 20 jurisdictions including Kenya, Nigeria and South Africa.

Across these campaigns the pattern is consistent. Humans choose targets, establish initial access and make strategic decisions; agents perform the high-volume work in between. Defence therefore has two aims: deny the human the foothold, and detect the agent once it is working.

## 02 Pillar 1: Deny the foothold

Every agentic campaign Unit 42 observed began with a web shell planted through a vulnerable web application. The initial access is ordinary; the scale and speed that follow are not.

- Maintain a continuous inventory of internet-facing applications and remove anything without an owner.
- Patch internet-facing applications on a timeline measured in days, prioritising known exploited vulnerabilities.
- Restrict file-upload functionality: validate file type and content, store uploads outside the web root, and prevent execution in upload directories.
- Place administrative and developer platforms (CMS, GitLab, CI/CD, cloud consoles) behind multi-factor authentication and IP restrictions.

## 03 Pillar 2: Detect the agent

Agentic intrusions behave differently from human operators. Unit 42 measured a median gap between commands of 1.4 seconds under agent control, against 23.9 seconds under human control. Agents also organise output with structured delimiters, leave reasoning comments in scripts, and generate dozens of unique encoded payloads per session. These traits support specific detections.

- Alert on interactive sessions sustaining sub-two-second command rates on servers.
- Alert on enumeration of many cloud services or regions in a short period from a single identity.
- Inspect scripts for structured delimiters and natural-language reasoning comments.
- Alert on bursts of unique Base64-encoded PowerShell or Python executions.
- Monitor cloud instance-metadata access and enforce IMDSv2.

## 04 Pillar 3: Respond in minutes

An agent that can move from web shell to cloud reconnaissance in 140 minutes will finish its work before a next-morning alert review. Organisations need either 24-hour monitoring with authority to contain, or automated containment for high-confidence detections, such as isolating a host or revoking a session, with human review immediately afterwards. Pre-authorised response actions, agreed in writing, prevent delay while approvals are sought.

## 05 Pillar 4: Govern your own agents

Organisations are deploying AI agents that read email, query databases, raise tickets and take actions. Each is effectively a privileged account that can be manipulated through the content it processes. Prompt injection, meaning instructions hidden in a document, web page or email that the agent then follows, is the defining risk.

- Grant each agent the minimum permissions needed, scoped per task, and never the credentials of a human administrator.
- Require human approval for consequential actions: payments, deletions, external communications and permission changes.
- Log every tool call and action an agent takes, with retention long enough for investigation.
- Treat all content an agent reads as untrusted input, and test agents against prompt-injection scenarios before deployment.
- Protect AI service API keys: keep them out of code and front-end bundles, scope and rotate them, and set billing alerts to detect token jacking.

## 06 A 90-day plan

The framework can be implemented incrementally. The following sequence delivers the greatest risk reduction first.

- Days 1–30: inventory internet-facing applications, close or patch high-risk exposures, enforce MFA on administrative and developer platforms.
- Days 31–60: deploy behavioural detections for machine-speed activity and cloud enumeration; agree pre-authorised containment actions.
- Days 61–90: inventory internal AI agents and integrations, apply least privilege and approval gates, and run a prompt-injection test.

## References

1. Unit 42: Recent trends and tactics of agentic AI attacks (November 2026)
   https://unit42.paloaltonetworks.com/autonomous-ai-agent-threat-campaigns/
2. Unit 42: Cairn framework used in autonomous attacks (November 2026)
   https://unit42.paloaltonetworks.com/cairn-framework-ai-autonomous-attacks/
3. Zscaler ThreatLabz: Inside a global AI-assisted hacking campaign (October 2026)
   https://zscaler.com/blogs/security-research/inside-global-ai-assisted-hacking-campaign
4. Unit 42: AI insights from the 2026 Global Incident Response Report (July 2026)
   https://unit42.paloaltonetworks.com/ai-insights-incident-response-report/

## About this publication

This document is published by SCIAT AFRICA, a cybersecurity and digital forensics agency based in Nairobi, Kenya. It is free to read, print and circulate inside your organisation. The web edition, which carries the same content, is at https://sciatafrica.com/research/defending-at-machine-speed-agentic-ai.

It is general guidance, not legal advice and not a substitute for an assessment of your own environment. Where it refers to a law, regulation or standard, read the instrument itself before relying on it — and talk to us if you want help applying it to your estate.

**SCIAT AFRICA**

info@sciatafrica.com · +254 794 946633 · https://sciatafrica.com