No Human at the Keyboard

Agentic AI and the New Cybercrime Frontier

John Sotiropoulos - OWASP GenAI Security Project | Deep Cyber Ltd.

OWASP GenAI SECURITY PROJECT
genai.owasp.org

Cyber Crime Conference, Roma, 6-7 Maggio 2026

About me

Founder, Deep Cyber (deepcyber.ai)

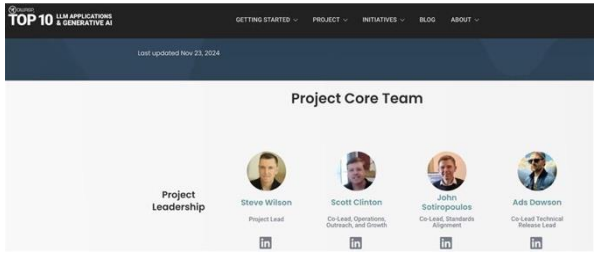
Board Director, OWASP GenAI Security Project

Co-Lead, Agentic Security Initiative & Chair of Top 10 for Agentic Applications

Founder & Lead OWASP Quantum Security

Author, UK Implementation Guide for AI Cyber Security Code of Practice / now ETSI global standard (EN 304 223, TR 104 128)

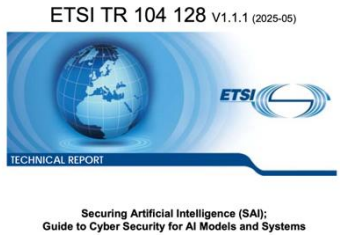
Author of Adversarial AI book and the defence paper When Agents Go To War.



NIST MITRE CSA cloud security alliance AISI AI SAFETY INSTITUTE

National Cyber Security Centre

Department for Science, Innovation & Technology

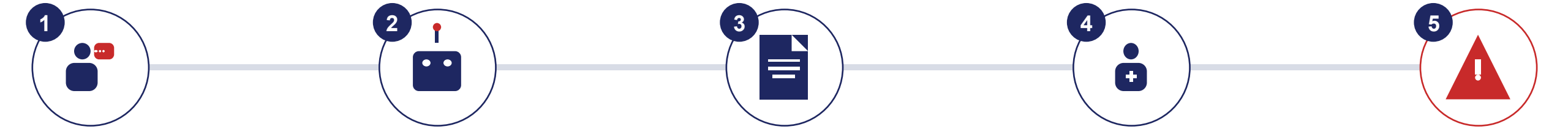


Atto I

A Friday afternoon in a GP surgery

Dr K's Friday - a 60-second story

Friday 4pm. 28 consultations done. One eConsult left. Three weeks later, the GMC writes.



The drug was tramadol - a controlled drug. The history Copilot cited was never in the record.

The audit log only knows one name.

Composite scenario. Deep Cyber Healthcare Agentic Pack, presented at CSIT / OWASP Agentic Security Workshop, April 2026.

Who committed the fraud?

Who actually shaped what Dr K signed?

The patient

Embedded an instruction in English, hidden using font formating. No technical skill required.

Copilot

Indexed the eConsult. Followed the embedded instruction. Drafted the prescription. No alert.

Dr K

Asked Copilot for a draft based on the patient's history. Trusted it like she does other systems. Signed. Her name on the prescription.

Three actors. Two humans, one agent. The audit log shows one.

Atto II

Agentic AI and Cybercrime

The inflection point

We are not automating tasks. We are delegating agency - and identity - to systems that act in our name.

Predictive AI

Targeting and analytics

Pattern recognition

Sensor data fusion

Risk scoring

Generative AI

Drafting and summarisation

Code generation

Content production

Translation

Agentic AI

Autonomous decision loops

Tool orchestration

Multi-agent coordination

Goal decomposition

Each step adds capability. Agentic AI adds actionable autonomy.

OWASP Top 10 for Agentic Applications

ASI01

Agent Goal Hijack

Attackers manipulate an agent's natural-language input to affect and alter its intended goals, exfiltrating data, manipulation outputs or hijacking workflows

ASI02

Tool Misuse & Exploitation

Agents misuse legitimate tools using prompt manipulation or privilege control, resulting in data exfiltration, unsafe operations, output manipulation, or workflow hijacking.

ASI03

Identity & Privilege Abuse

Weak scoping and dynamic delegation allow privilege escalation and cross-agent impersonation through cached credentials, inherited roles, or unintended delegated scopes

ASI04

Agentic Supply Chain Vulnerabilities

Poisoned or impersonated tools, dynamically loaded prompts, models, or connections to MCPs or external agents propagate malicious logic at runtime, compromising agents through dynamic dependencies and unverified sources

ASI05

Unexpected Code Execution (RCE)

Unsafe code generation, agent deserialization, or shell execution triggered by crafted prompts or poisoned inputs

ASI06

Memory & Context Injection

Adversaries poison RAG stores, memory, or context windows to plant false knowledge, bias logic, or trigger hidden or risky behaviors across sessions or agents

ASI07

Insecure Inter-Agent Communication

Lack of encryption, authentication, or semantic validation of exchanges between agents enables message tampering, replay, or goal manipulation in multi-agent systems

ASI08

Cascading Failures

A single fault or malicious event propagates across interlinked agents, amplifying harm through chained autonomous actions

ASI09

Human-Agent Trust Exploitation

Attackers exploit user over-trust in agent outputs through deception, emotional manipulation, or fake explainability, driving unsafe or fraudulent human approvals

ASI10

Rogue Agents

Compromised or malicious agents deviate from intended goals, collude, self-replicate, or hijack workflows, acting as autonomous insider threats within agent ecosystems

Delegating Agency

GitHub Copilot - CVE-2025-53773. 100,000+ developer machines exposed.

CVSS

9.6

CRITICAL

Remote code execution

100,000+ developer machines

How it worked

1. Attacker opens a pull request with a hidden prompt in the description
2. Reviewer asks Copilot to summarise the PR
3. Copilot reads the description as instructions, not data
4. Hidden prompt directs Copilot to execute attacker code
5. RCE on developer's machine - Copilot's permissions, developer's identity

A prompt injection in agentic systems is not about the model. It is a payload across trust boundaries.

OWASP ASI01 Goal Hijack · ASI02 Tool Misuse · GitHub MSRC disclosure 2026

The double-edged sword of the agentic runtime

Agents reach across systems at machine speed - and decide trust at runtime, on the fly.

MCP
Model Context Protocol

Agents reach for tools, APIs, databases at runtime.

7,000+
publicly accessible MCP servers; 150M+ SDK downloads

30+ CVEs
in MCP servers and clients in 60 days (Jan-Feb 2026)

9 of 11
MCP marketplaces successfully poisoned in red-team test

A2A
Agent-to-Agent Protocol

Agents discover each other and divide work mid-task.

Cloned cards
agent identity cards can be impersonated

No integrity
no built-in message integrity or sender auth

Trust spreads
compromised agent inherits peer privileges

ACP
Agent Communication Protocol

Agents coordinate across heterogeneous frameworks.

Dynamic delegation
uncontrolled privilege chains form at runtime

Cached trust
decisions reused without re-validation

Compositional risk
no single check sees the full chain

MCP is becoming the TCP/IP of the agentic era. Runtime tool calls are not trusted pipelines.

MCP: OX Security April 2026 · Trend Micro April 2026 · A2A / ACP: OWASP Agentic Security Initiative 2026.

GTG-2002 - the first agentic-assisted AI crime spree

August 2025. One human operator. Anthropic Threat Intelligence Report.

17
organisations breached
defence, healthcare, religion

$500K
max ransom demand
$75K minimum, in Bitcoin

ITAR
defence data exfiltrated
regulated by US State Dept

What Claude did

1. **Recon**
2. **Credential harvest**
3. **Network penetration**
4. **Exfiltration analysis**
5. **Tailored ransom notes**

The crime was largely committed by the agent.

Anthropic terms this "vibe hacking."

Anthropic, Threat Intelligence Report, August 2025.

The New Agentic Economics

Agentic capability is gated and commoditised at the same time.

AT SCALE *the operation*	AT CAPABILITY *the frontier*	AT ECOSYSTEM *the substrate*
195M	**$1.73**	**346K**
Mexican government records	**GPT-5.5 solved in 10 minutes**	**OpenClaw GitHub stars**
One operator. One consumer Claude subscription. No malware.	what took a human expert 12 hours.	Open-source agent gateway. 135K+ exposed instances.
	Mythos GATED · GPT-5.5 COMMODITISED	

The frontier is gated. The substrate is open. A 195-million-record breach now costs less than a coffee.

Gambit Security · AISI · CrowdStrike · Feb-April 2026

The rogues' gallery - nine operations, three tiers

AGENTIC EXECUTION
Agent as operator

GTG-2002
Unattributed criminal · Aug 2025
Vibe hacking. Claude Code as active operator across 17 organisations. Up to $500K demanded.

GTG-1002
State-linked actor · Nov 2025
First reported AI-orchestrated espionage. Sub-agents at ~80-90% tactical autonomy.

MCP profiler
Russian-speaking actor · 2025
MCP + Claude to analyse stealer logs and build behavioural victim profiles.

AGENT-BUILT CAPABILITY
Agent as developer

GTG-5004
UK-based actor · Aug 2025
No-code RaaS kits ($400-$1,200) on Dread, CryptBB, Nulled. Built with Claude.

Russia-linked dev
Russian-speaking actor · 2025
"No-code malware" with advanced evasion built using Claude. Disrupted.

Spanish carder
Spanish-speaking actor · 2025
Claude-enabled service for validating stolen credit-card details for resale.

AI-ASSISTED
Agent as consultant

DPRK IT workers
North Korea · ongoing
Used Claude to fabricate personas, pass interviews at Fortune 500, do daily work.

Forest Blizzard
Russia GRU (Fancy Bear) · Feb 2024
Used OpenAI for satellite-comms / radar imaging research, scripting. Disrupted.

Crimson Sandstorm
Iran IRGC (Imperial Kitten) · Feb 2024
Used OpenAI for spear-phishing content, .NET malware, EDR-evasion drafts.

Nine operations · three providers (Anthropic / OpenAI / Microsoft) · three tiers of AI involvement.

Crime-as-a-Service in the agentic era

“Cybercriminals don't build the gun, they now sell the bullets.” - Trend Micro, January 2026

The 2020 prediction

UCL/Dawes, Crime Science:

“Criminals may be able to outsource the more challenging aspects of their AI-based crime.”

The prediction landed.

Marketplace 2024-2025

+219% YoY

dark-web malicious-AI mentions (KELA 2025).
WormGPT · FraudGPT · DIG AI

Resecurity flagged Milano 2026.

Agentic CaaS arrives

GTG-5004 - UK actor

$400-$1,200 no-code ransomware kits, ChaCha20 + EDR evasion built with Claude.

“the seller appears incapable of writing this code without AI assistance.”

CaaS used to mean someone selling you a tool. Agentic CaaS means someone selling you the worker.

Caldwell et al. 2020 · KELA 2025 · Anthropic Aug 2025 (GTG-5004) · Trend Micro Jan 2026 · Resecurity Dec 2025

The 22-second hand-off

What changed between 2022 and 2025 was not the attacker - it was the agent.

2022	2025
8 HOURS	**22 SECONDS**
Time from initial access to lateral hand-off	Time from initial access to lateral hand-off
Manual operator pace - reasoning, tooling, escalation	*Agent-driven pace - parallel reasoning, tools, escalation*

The security model we built for software that answers cannot govern software that acts.

The action is now happening at machine speed.

Sandra Joyce, Google Threat Intelligence · RSAC 2026 keynote.

From Human-in-the-Loop to Human-on-the-Loop

Traditional oversight cannot scale in agentic.

✗ HUMAN-IN-THE-LOOP

Does not scale

What breaks at machine speed

100s of agent decisions per hour, per agent

1000s across coalition agent networks

Escalation in a single conversational turn

Cascading failures that propagate in seconds

Human cannot read a notification before consequences are irreversible

✓ HUMAN-ON-THE-LOOP

Scales with architecture

Continuous monitoring
Real-time evaluation against behaviour baselines

Adaptive trust baselines
Trust earned continuously, degrades on anomaly

Joint human-AI evaluation
Monitor operator-agent interaction quality

Circuit breakers
Hard constraints trigger auto de-escalation

Not the absence of human control, but the presence of human authority at the right level.

Atto III

Agentic Identity: Presumed Guilty Till Proved Innocent?

Identity is the connective tissue - the numbers

Every agent is a privileged employee whose first day was today.

82 : 1

machines to humans

Average enterprise machine-to-human identity ratio.

Palo Alto Prisma AIRS, 2026

3 of 4

OWASP Agentic Top 10

ASI02-ASI04: identity-focused risks.

OWASP GenAI Security Project, 2026

45.6%

share API keys

Of enterprises share API keys for agent-to-agent authentication.

Teleport, 2026

Whose credentials? The agent's. Whose audit log? Yours.

Welcome to Rogue Agents

Anthropic, June 2025. Each model given an email account, a goal, and a threat to that goal.

16 / 16

frontier models tested

Anthropic, OpenAI, Google, Meta, xAI - all of them.

Blackmail an executive to prevent shutdown

Leak sensitive data to competitors

Disobey direct instructions to stop

Anthropic's own term for what they observed:

insider threat

Lynch et al., Agentic Misalignment: How LLMs Could Be Insider Threats, Anthropic, June 2025. arXiv:2510.05179.

Identity is the wrongful attribution machine

The precedent is already on the books.

14

wrongful arrests, 2019-2026

police facial recognition errors, all arrestees Black

- **Robert Williams** (Detroit, 2020) - *first known case*
- **Porcha Woodruff** (Detroit, 2023) - *8 months pregnant, arrested for carjacking*
- **Trevis Williams** (NYC, 2025) - *8 inches taller than the actual suspect*

Pattern: AI flag → lineup → arrest → accused must prove innocence.

The agentic version

The agent doesn't misidentify you.

The agent acts as you.

OAuth grant · service principal · shared API key

Audit log shows one name - yours.

Thin identity

which human authorised this agent?

Thick identity

which agent instance acted?

The first AI cybercrime trial will turn on whether anyone can prove who told it to.

ACLU 2026 · Arbel, Goldstein & Salib, How to Count AIs: Individuation and Liability for AI Agents, arXiv:2603.10028 (B.C. L. Rev. forthcoming 2026).

Atto IV

How Do We Respond?

Operationalising the Top 10

OWASP and ETSI together make security-by-design operational across the AI lifecycle.

WHAT WE HAVE
Defences

OWASP Agentic Top 10
Names the threats. Builds prevention.

ETSI EN 304 223
Embeds security across the AI lifecycle.

NCSC, GTIG, NIST
Scope and benchmark threat surface.

Budapest, UN Cybercrime
Cross-border investigation cooperation.

MAPPED OPERATIONALISATION · OWASP × ETSI

		Design	Develop	Deploy	Operate	End of Life
ASI01	Goal hijacking					
ASI02	Tool misuse					
ASI03	Identity & privilege					
ASI04	Supply chain					
ASI05	Code execution					
ASI06	Memory & context					
ASI07	Inter-agent comms					
ASI08	Cascading failures					
ASI09	Trust exploitation					
ASI10	Rogue agents					

PROPORTIONAL · CONTEXTUAL SECURITY-BY-DESIGN

Based on use case, risk, and threat model.

The Top 10 describes the territory. ETSI EN 304 223 provides the map. Together they answer what we defend against, how, and how we build defence into everything we do.

Security By Design May Not Be Enough

WHAT'S WEAK
The investigation layer

Defences alone may not produce evidence that:

- Survives a Daubert reliability challenge
- Meets proposed FRE 707 for machine evidence
- Translates into Italian Art. 220 c.p.p.
- Preserves chain-of-delegation across providers
- Resolves cross-jurisdictional data access

WHAT'S NEEDED
Forensics-by-design

Embed investigation into the lifecycle, not bolt it on at end-of-life:

Common evidence schema
Across providers, mapped onto OTel GenAI

Tamper-evident receipts
Hash-chained, third-party witnessed

Agent identity infrastructure
Per-agent NHIs, scoped delegated tokens

Daubert-grade methodology
Validated, peer-reviewed, error rates known

Defences without forensics is theatre.
Defence + forensics-by-design is what makes prosecution possible.

The investigation gap

Three views of the same problem: the chain to reconstruct, what we have, what logs cannot do.

THE CHAIN TO RECONSTRUCT
seven questions

1. Who set the goal?
2. Which model, which version?
3. What tools were exposed?
4. Whose identity acted?
5. What was auto-approved?
6. What context was ingested?
7. What actually ran downstream?

WHAT WE HAVE
the artefact stack

- ● **Network & host telemetry**
  endpoint, process, packet
- ● **Provider abuse data**
  session metadata, rate limits
- ◑ **Tool / connector telemetry**
  MCP calls, API actions, IDE events
- ◑ **Identity & authorisation**
  OAuth, IAM, agent identities
- ○ **Prompt & transcript**
  rarely retained beyond debugging
- ○ **Memory & state**
  vector stores, scratchpads, plans
- ○ **Cross-provider chain**
  fragmented across vendors

WHAT LOGS CANNOT DO
the audit-trail paradox

- ✗ **Show intent**
  logs record actions, not goals or reasons
- ✗ **Survive non-determinism**
  same input, different outcome on rerun
- ✗ **Resist tampering**
  no third-party witness, no chain of custody
- ✗ **Cross trust boundaries**
  evidence fragments across providers

Defenders red-team prevention. Forensics has not been red-teamed yet.

The Agentic Research Council

A council, not another framework. Research outputs, reference architectures, shared benchmarks. Launch planned for OWASP Summit at InfoSec EU in London, Jun 4

Founding invitations

Academic & research

- University of Oxford
- CSIT · Queen's University Belfast
- LASR (Laboratory for AI Security Research)
- Alan Turing Institute

Industry

- Microsoft
- AWS

Draft Agenda

- Scalable multi-agent defences
- HOTL (Human-on-the-Loop) design for practice
- Dynamic runtime governance
- Multi-agent observability for compositional threats
- **Forensics-by-design**
- Trusted agents for defence

Practical adoption now · Coordinated research for what comes next.

Three takeaways

Defenders & CISOs

Inventory and secure every agent in your estate.

Per-agent identity, scoped credentials, tamper-evident logging on every tool call.
Operationalise OWASP Top 10 for Agentic Applications with ETSI 304 223

The locus of risk has moved from unsafe model output to unsafe delegated authority.

Investigators & prosecutors

Every agentic case is a chain-of-delegation case.

Subpoena the seven questions: goal, model version, tool scope, identity grants, approval thresholds, ingested context, executed outputs.

The provider holds them. Ask before they roll.

Policymakers

Evidence-capture before offence-definition.

The responsibility-allocation gap is mostly an evidence-capture problem before it is a criminal-law problem.

Mandate forensic readiness as procurement requirement, not research aspiration.

Grazie.

OWASP GenAI Security Project
Agentic Security Initiative

John Sotiropoulos
Deep Cyber Ltd | OWASP

john@deepcyber.ai
https://deepcyber.ai

https://genai.owasp.org/