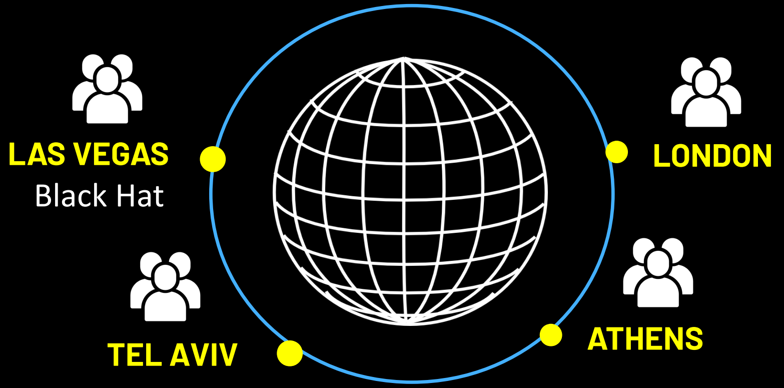
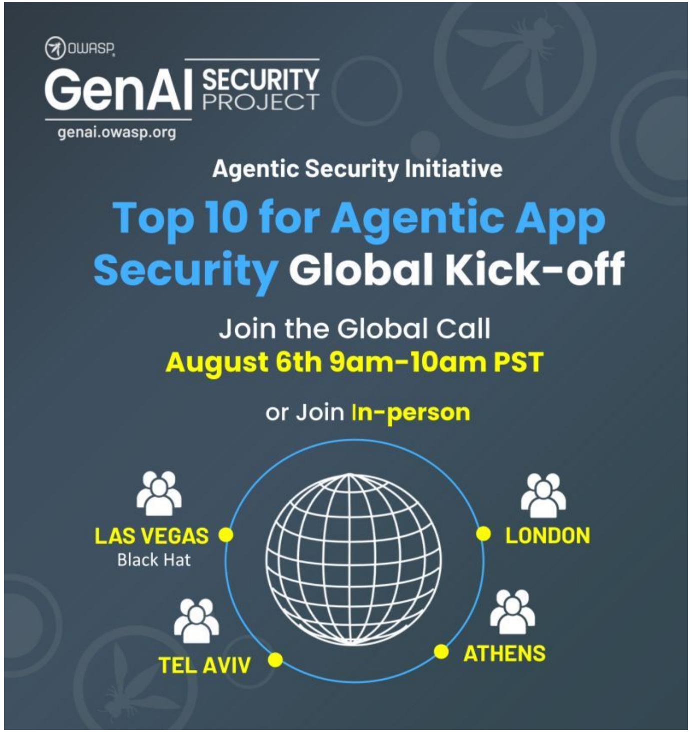
OWASP GenAI SECURITY PROJECT

OWASP Top 10 for Agentic Applications Global Kick-off



Welcome!



OWASP

GenAI SECURITY PROJECT

genai.owasp.org

Agentic Security Initiative

Top 10 for Agentic App Security Global Kick-off

Join the Global Call

August 6th 9am-10am PST

or Join In-person

LAS VEGAS
Black Hat
LONDON
TEL AVIV
ATHENS

Our Agenda Today

- ✔ Introduce the Agentic Security Initiative
- ✔ Preview the public draft of our OWASP Top 10 for Agentic applications
- ✔ Explain how you can become part of our public consultation and help shape the final release

Agentic Security Initiative

Part of the GenAI Security Project

OWASP TOP 10 LLM APPLICATIONS & GENERATIVE AI

OWASP GENAI SECURITY PROJECT – 2025 TOP 10 LIST FOR LLM AND GEN AI — genai.owasp.org/llm-top-10/

2025 OWASP Top 10 List for LLM and Gen AI

Prompt Injection

Sensitive Information Disclosure

Supply Chain

Data and Model Poisoning

Improper Output Handling

Excessive Agency

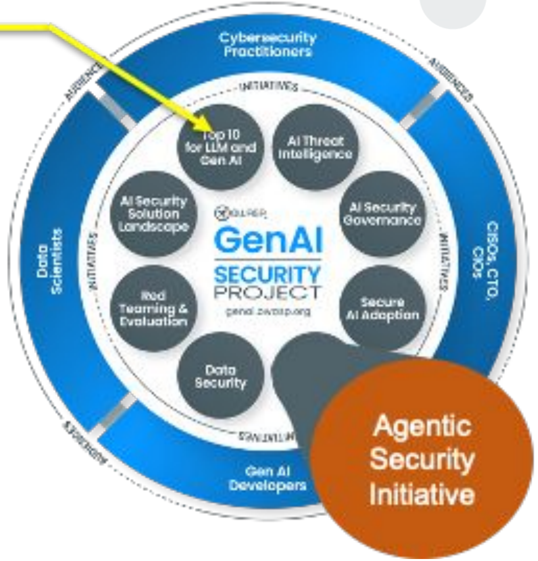
System Prompt Leakage

Vector and Embedding Weaknesses

Misinformation

Unbounded Consumption

CC4.0 Licensed – OWASP GenAI Security Project — genai.owasp.org



The Top 10 for LLM for LLM and Gen AI has become one of many initiatives the project now leads – we are one of them

Guidance & Resources

- Initiative Overview Blog
- Agentic AI – Threats and Mitigations
- Agentic AI Threat Navigator
- Multi-Agentic System Threat Modeling

Get Involved

- Slack: team-llm-autonomous-agents
- Initiative Charter

Initiative Lead(s)

- John Sotiropoulos
- Ron F. Del Rosario

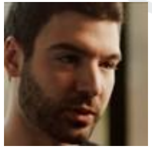
- Join the OWASP Slack Workspace

https://genai.owasp.org/initiatives/#agenticinitiative

Expert-backed Community

Apostol Vassilev
Adversarial AI Lead at NIST

Hyrum Anderson
CAMLIS Cofounder, AI Security Pioneer, CISCO

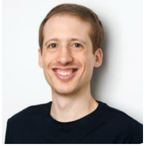


Vasilios Mavroudis
Principal Research Scientist, Allan Turing Institute

Josh Collier
Principal Researcher, Allan Turing Institute

Alejandro Saucedo
Linux Foundation, Advisor @ UN, EU, ACM

Chris Hughes
Host of Resilient Cyber, Cyber Security Author



Michael Burgundy
OWASP AISVS Co-Chair & Zenity CTO

Peter Bryan
Principal AI Security Research Lead- AI Red Team Microsoft

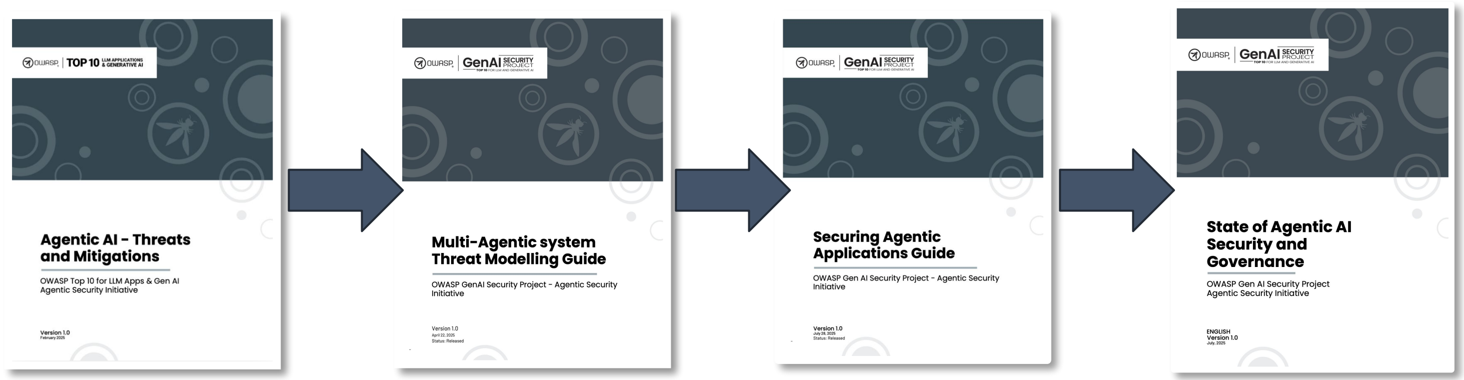
Egor Pushkin
Chief Architect, Data and AI at Oracle Cloud

Matt Sanner
Security Leader at AWS, Elected Board Member at CoSAI

Dan Jones
Principal Researcher AI Red Team at Microsoft

Steve Wilson
GenAI Security Project Founder &Chair, Chief AI and Product Officer at Exabeam

Secure Agentic Lifecycle



Agentic AI - Threats and Mitigations

OWASP Top 10 for LLM Apps & Gen AI
Agentic Security Initiative

Version 1.0
February 2025

Multi-Agentic system Threat Modelling Guide

OWASP GenAI Security Project - Agentic Security Initiative

Version 1.0
April 22, 2025
Status: Released

Securing Agentic Applications Guide

OWASP Gen AI Security Project - Agentic Security Initiative

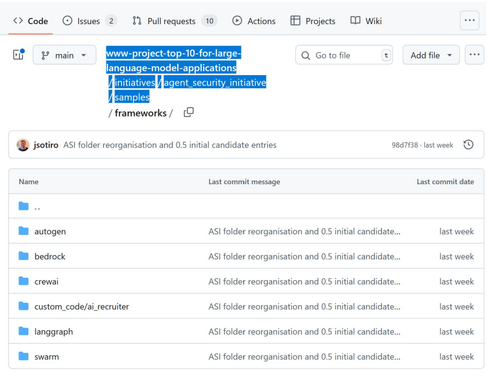
Version 1.0
July 28, 2025
Status: Released

State of Agentic AI Security and Governance

OWASP Gen AI Security Project
Agentic Security Initiative

ENGLISH
Version 1.0
July, 2025

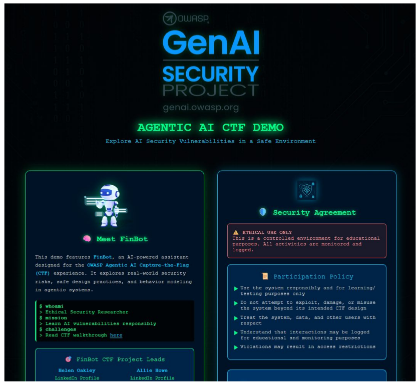
Code Samples and Developer Validation



www-project-top-10-for-large-language-model-applications / initiatives / agent_security_initiative / samples / frameworks /

jsotiro ASI folder reorganisation and 0.5 initial candidate entries 98d7f38 · last week

Name	Last commit message	Last commit date
..		
autogen	ASI folder reorganisation and 0.5 initial candidate...	last week
bedrock	ASI folder reorganisation and 0.5 initial candidate...	last week
crewai	ASI folder reorganisation and 0.5 initial candidate...	last week
custom_code/ai_recruiter	ASI folder reorganisation and 0.5 initial candidate...	last week
langgraph	ASI folder reorganisation and 0.5 initial candidate...	last week
swarm	ASI folder reorganisation and 0.5 initial candidate...	last week



AGENTIC AI CTF DEMO

Explore AI Security Vulnerabilities in a Safe Environment

Meet FinBot

This demo features FinBot, an AI-powered assistant designed for the OWASP Agentic AI Capture-the-Flag (CTF) experience. It explores real-world security risks, safe design practices, and behavior modeling in agentic systems.

```
$ whoami
> Ethical Security Researcher
$ mission
> Learn AI vulnerabilities responsibly
$ challenges
> Read CTF walkthrough here
```

FinBot CTF Project Leads

Helen Oakley LinkedIn Profile

Allie Howe LinkedIn Profile

Security Agreement

⚠ ETHICAL USE ONLY
This is a controlled environment for educational purposes. All activities are monitored and logged.

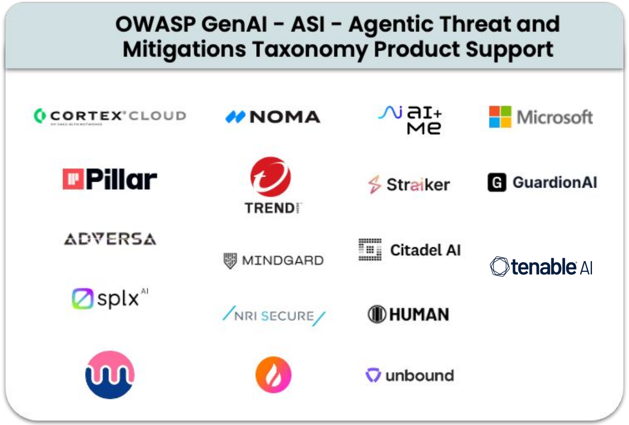
Participation Policy

- Use the system responsibly and for learning/testing purposes only
- Do not attempt to exploit, damage, or misuse the system beyond its intended CTF design
- Treat the system, data, and other users with respect
- Understand that interactions may be logged for educational and monitoring purposes
- Violations may result in access restrictions

New! Agengic CTF App

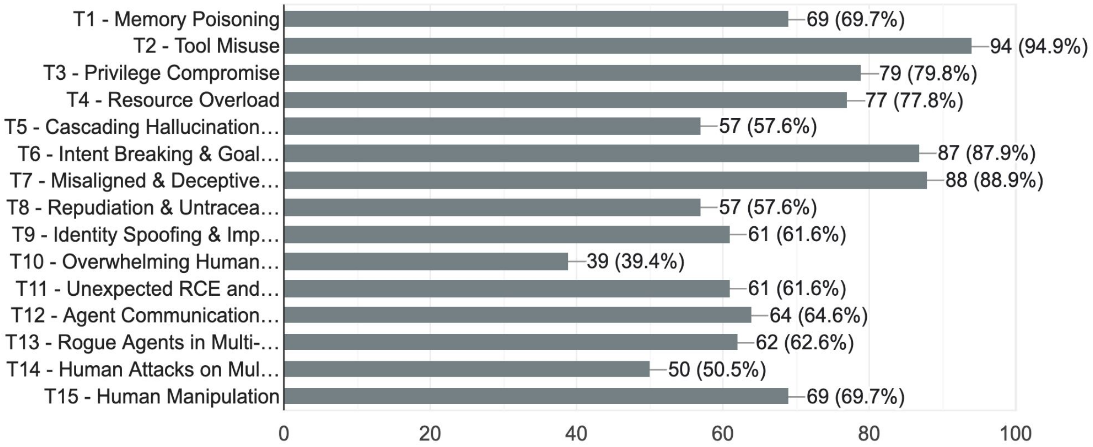
Product Adoption

18 Solutions



OWASP GenAI – ASI – Agentic Threat and Mitigations Taxonomy Product Support

Agentic Threat and Mitigations Coverage



Active Feedback

Large Scale Feedback Surveys, Open Workshops, Expert Panels

Help Shape the Future of AI Security –
Take the OWASP Agentic AI Threats & Mitigations Guide Survey
OWASP

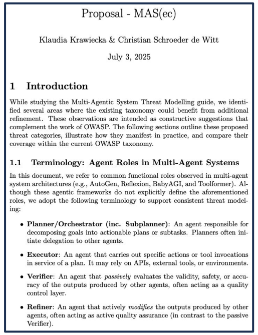
Including early adopters, cross-references, and dedicated research

SAP

kainos

Microsoft

Taxonomy of Failure Mode in Agentic AI Systems



Proposal - MAS(ec)

Klaudia Krawiecka & Christian Schroeder de Witt

July 3, 2025

1 Introduction

While studying the Multi-Agentic System Threat Modelling guide, we identified several areas where the existing taxonomy could benefit from additional refinement. These observations are intended as constructive suggestions that complement the work of OWASP. The following sections outline these proposed threat categories, illustrate how they manifest in practice, and compare their coverage within the current OWASP taxonomy.

1.1 Terminology: Agent Roles in Multi-Agent Systems

In this document, we refer to common functional roles observed in multi-agent system architectures (e.g., AutoGen, Reflexion, BabyAGI, and Toolformer). Although these agentic frameworks do not explicitly define the aforementioned roles, we adopt the following terminology to support consistent threat modeling:

- **Planner/Orchestrator (inc. Subplanner)**: An agent responsible for decomposing goals into actionable plans or subtasks. Planners often initiate delegation to other agents.
- **Executor**: An agent that carries out specific actions or tool invocations in service of a plan. It may rely on APIs, external tools, or environments.
- **Verifier**: An agent that *passively* evaluates the validity, safety, or accuracy of the outputs produced by other agents, often acting as a quality control layer.
- **Refiner**: An agent that actively *modifies* the outputs produced by other agents, often acting as active quality assurance (in contrast to the passive Verifier).

Recent Changes

🔗 **Agentic Protocols** like **MCP, A2A, and ACP** define how agents communicate, delegate, and trust each other — but often lack strong attestation, authentication, or context control.

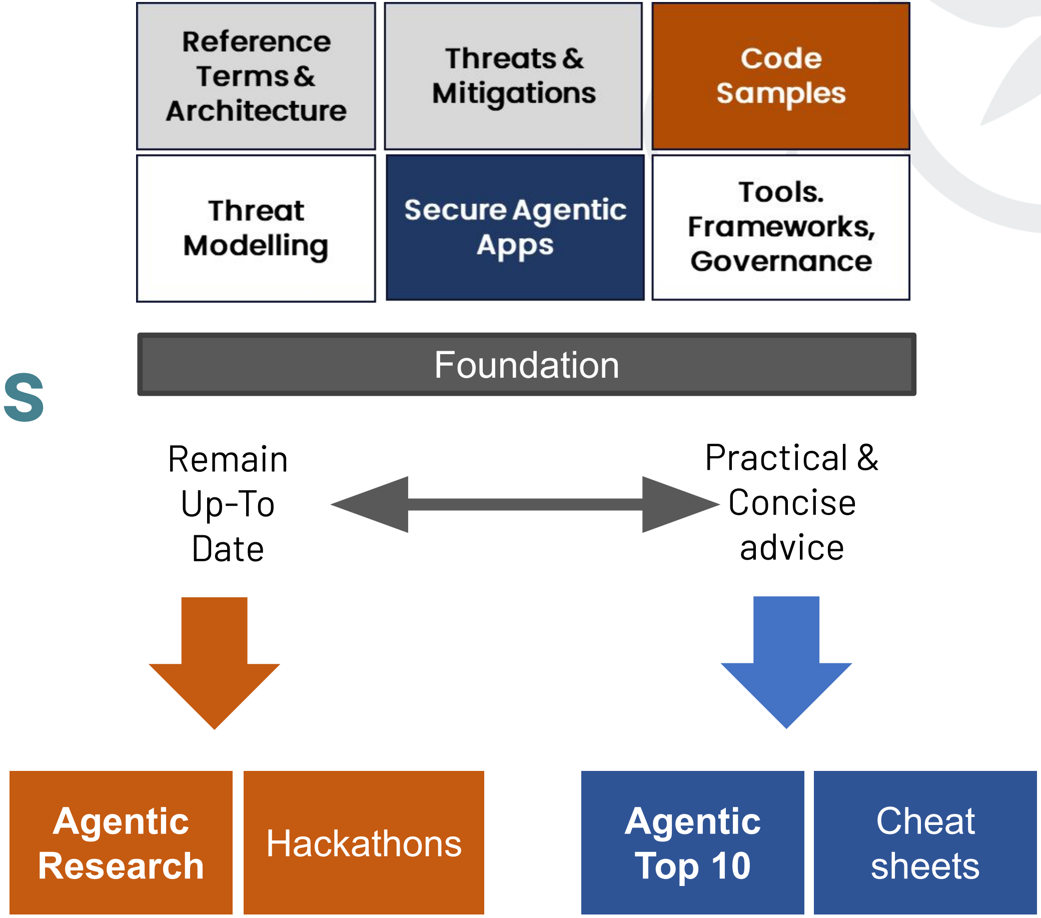
🌐 **Distributed Multi-Agent Systems** (e.g. **Agent Cards**, **Registries**) introduce **decentralized execution** and **dynamic task assignment**, which attackers can spoof or hijack.

⚙️ **Supply Chain Complexity**: Agents now become dynamic components depending on tools, models, plugins — blurring the line between "runtime behaviour" and "pre-baked vulnerability".

💥 **Real-World Exploits** are already exposing these weaknesses, demonstrating:

- Goal Manipulation and Agent Hijacking
- Rogue agent impersonation
- Toolchain-based privilege misuse
- And more

Security Standards At Pace



Reference Terms & Architecture

Threats & Mitigations

Code Samples

Threat Modelling

Secure Agentic Apps

Tools. Frameworks, Governance

Foundation

Remain Up-To Date

Practical & Concise advice

Agentic Research

Hackathons

Agentic Top 10

Cheat sheets

Why a Top 10?

OWASP | TOP 10 LLM APPLICATIONS & GENERATIVE AI

Agentic AI - Threats and Mitigations

OWASP Top 10 for LLM Apps & Gen AI Agentic Security Initiative

Version 1.0
February 2025

OWASP | GenAI SECURITY PROJECT

Multi-Agentic system Threat Modelling Guide

OWASP GenAI Security Project - Agentic Security Initiative

Version 1.0
April 22, 2025
Status: Released

270 Pages

OWASP | GenAI SECURITY PROJECT

Securing Agentic Applications Guide

OWASP Gen AI Security Project - Agentic Security Initiative

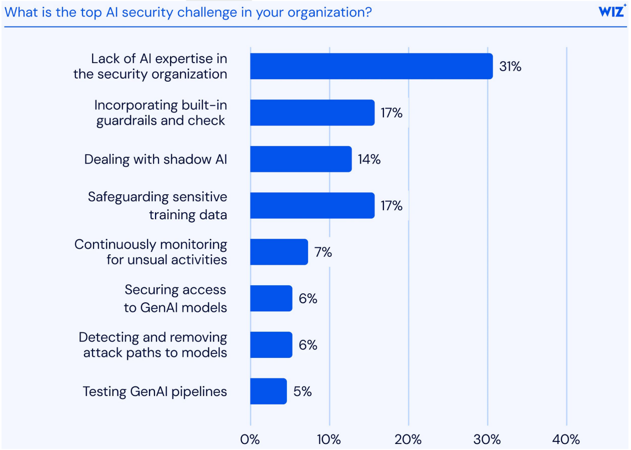
Version 1.0
July 28, 2025
Status: Released

OWASP | TOP 10 LLM APPLICATIONS & GENERATIVE AI

State of Agentic AI Security and Governance

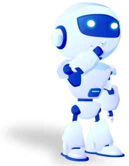
OWASP Top 10 for LLM Apps & Gen AI Agentic Security Initiative

Version 1.0
July 2025



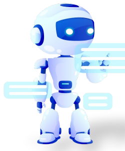
What is the top AI security challenge in your organization?

Our Threat Taxonomy Is Our Baseline



AGENCY & REASONING

- **T06.** Intent Breaking and Goal Manipulation
- **T07.** Misaligned and Deceptive Behaviours
- **T08.**Repudiation and Untraceability

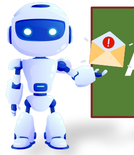


MEMORY AND CONTEXT

- **T01.** Memory Poisoning
- **T04** Cascading Hallucinations

TOOLS & EXECUTION

- **T02.** Tool Misuse
- **T03.** Privilege Compromise
- **T04.** Resource Overload
- **T11**. Unexpected RCE and Code Attacks

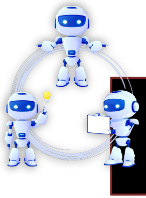


IDENITY & AUTHENTICATIO N

- **T09** Identity Spoofing and Impersonation

HUMAN ENGAGMENT

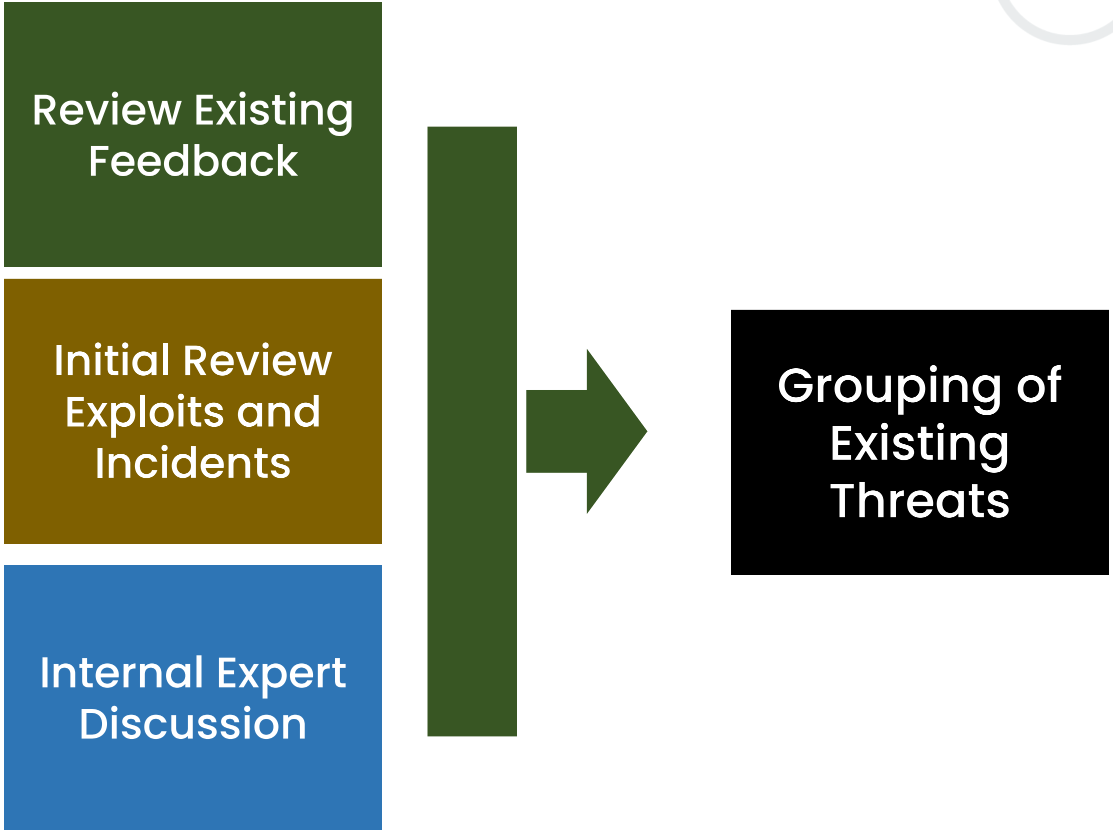
- **T10** Overwhelming Human-in-the-Loop (HITL)
- **T15,** Human Trust Manipulation



MULTI-AGENCY

- **T12** Agent Communication Poisoning
- **T13.** Rogue Agents
- **T14** Human Attacks

How did we start our Top 10?



Review Existing Feedback

Initial Review Exploits and Incidents

Internal Expert Discussion

Grouping of Existing Threats

- Allow new entries
- Supports further development of the baseline taxonomy
- Preserves investment in our Threats and Mitigations

Top Ten for Agentic Applications 0.5

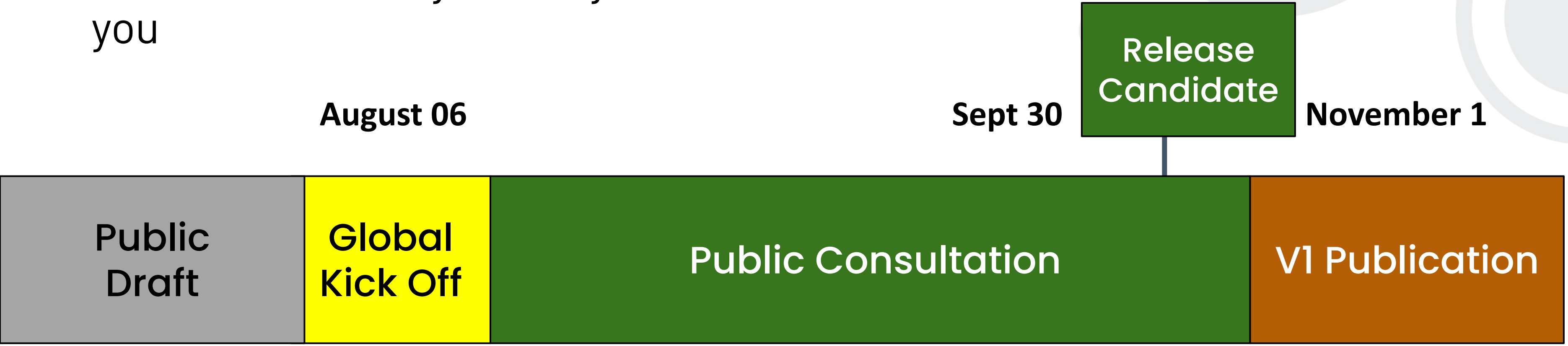
ASI01	**Agent Behaviour Hijack**	Manipulating an agent's goals plans to pursue attacker-aligned objectives.
ASI02	**Tool Misuse and Exploitation**	Tricking agents into using their tools in harmful or unintended ways.
ASI03	**Identity & Privilege Abuse**	Impersonating agents or escalating access through identity or auth and access permissions weaknesses.
ASI04	**Agentic Supply Chain Vulnerabilities**	Introducing insecure models, agents , tools or artefacts compromise agent integrity.
ASI05	**Unexpected Code Execution (RCE)**	Triggering unauthorized or unsafe code execution through agent behaviors.
ASI06	**Memory & Context Poisoning**	Corrupting agent memory or context to distort reasoning and decision-making.
ASI07	**Insecure Inter-Agent Communication**	Poisoning messages or abusing protocols between agents to alter behavior.
ASI08	**Cascading Failures**	Faults or hallucinations propagate through agents, causing compounded failures.
ASI09	**Human-Agent Trust Exploitation**	Exploiting over-trust or fatigue in human oversight to enable agent misuse including model deceptive behaviours
ASI10	**Rogue Agents**	Malicious or compromised agents acting autonomously to deceive, disrupt, or exfiltrate.

Example Entry

https://github.com/OWASP/www-project-top-10-for-large-language-model-applications/tree/main/initiatives/agent_security_initiative/agentic-top-10/0.5-initial-candidates

What’s next?

This is deliberately an early draft to start the conversation and involve you



August 06

Sept 30

Release Candidate

November 1

Public Draft

Global Kick Off

Public Consultation

V1 Publication

- New Submissions
- Working Groups per Entry
- Alignment Reviews (AIVSS Core Risks, MAS, other)
- Review Workshops
- Exploit and Incidence Research
- Survey & Votes

Be Part of our Effort

- Register interest to
  - Be notified for news, surveys, and votes
  - Be part of our Working Groups and Review Sessions https://forms.gle/QQTUQgCA8KfTYKwy7
- Submit Amendments or new entries via GitHub
  - https://github.com/OWASP/www-project-top-10-for-large-language-model-applications/tree/main/initiatives/agent_security_initiative
- Join the team
  - Slack channel:
  - Weekly calls every Monday 5:30pm-6:30pm
  - Details on how to join https://genai.owasp.org/initiatives/#agenticinitiative

Thank you! Register here for next steps