OWASP Top 10 for LLMs - Agentic Security Research Initiative

Revision	Date	Authors	Description
V1	18/11/2024	John Sotiropoulos	The initial version of a proposal to the core team to formally start the work
	19/11/2024	Ron F. Del Rosa	Intentional Vulnerable Samples and other edits
	04/12/2024	Ron F. Del Rosa	Added key components of Agentic AI systems
	11/12/2024	Scott Clinton	Approved by the Project CPMT

Initiative Abstract

The Agentic Security Research Initiative explores the emerging security implications of agentic systems, particularly those utilizing advanced frameworks (e.g., LangGraph, AutoGPT, CrewAI) and novel capabilities like Llama 3's agentic features. Our initial discussion on a proposed vulnerability in v2 of the Top 10 led to the conclusion that we need to understand this further, focusing on the unique vulnerabilities autonomous agents bring. A viewpoint has already been published in the Solutions Guide highlighting areas we should research more.

In addition to the security risks posed by traditional AI and Generative AI systems such as those identified in the OWASP Top 10 for Large Language Model Applications, Agentic AI systems present unique risks due to the following key components/capabilities:

- Planning
    a. Subgoal and decomposition
        i. Agents can break down large tasks into smaller, manageable suboals to be able to handle complex tasks. (Chain of thought, CoT: Wei et all. 2022)
    b. Reflection and refinement
        i. Agents are capable of self-criticism and self-reflection from previous actions, learn from mistakes and improve future results. (Reason + Act, ReAct: Yao et al.2023)
- Memory
    a. Agents have the ability to acquire, store, retain and retrieve information
        i. Sensory memory
            1. Embedding representations for multi-model inputs (text, image, video)
        ii. Short-term memory
            1. Short and finite memory restricted to the context window length. Limited to the single conversational thread with an agent.
        iii. Long-term memory
            1. Enables an agent to retain and recall information over extended periods via external vector stores that can easily be accessed during query time.
- Tool Use

a. Agents can invoke tools to accomplish tasks. Agents will have access to built-in tools (also known as function calling) such as browsing the web, conducting complex mathematical calculations, and generating executable code in response to a user query. Agents can also access more advanced tools via external API calls.

This initiative will analyse threats, vulnerabilities, and the challenges of scaling agentic systems, focusing on patterns of interaction, degree of autonomy, and the implications of transitive adaptation (e.g., self-reflection by LLMs). Outputs aim to support secure deployments by clearly understanding the new challenges and alignment with work in the community and industry.

Initiative Goals

Create an authoritative review of the security implications of autonomous agents to provide a navigational compass and recommendations for safe use and alignment of other agentic security work. This will cover:

- **Agentic Systems Landscape Analysis:**
  - A brief reference model noting frameworks, patterns, and tools for agentic systems, including on mobile devices ((e.g., Gemini on Pixel9). ***This is not intended to be a comprehensive cataloguing exercise*** on its own **but a concise reference** to set the scene and help consumers of our guidelines have a context and understand the implications on autonomy and operational contexts across both current single-agent and emerging multi-agent architectures.
- **Threat Modeling and Taxonomy for Agentic AI:**
  - Identify common security misconfigurations in well-known Agentic AI frameworks (LangChain, LangGraph, CrewAI, Microsoft Autogen, etc.),
  - Create sample exploits and/or intentionally vulnerable agentic AI implementation.
  - Help readers visualise the types of problems that can occur in agentic systems, both intentional, such as through indirect prompt injections, and unintentional, such as if an LLM misunderstands intent and calls a tool that does the wrong thing. Many of the potential ways that things can go wrong aren't obvious on the surface, and any help we can provide that gives a model for thinking about what can go wrong would be extremely helpful.
  - Develop a threat model for agentic systems and investigate new attack surfaces introduced by agent memory and local model deployments. The intention is to combine the threat modelling work with the sample exploits and implementations to inform of and demonstrate threats and vulnerabilities in action, highlighting what is real with high likelihood versus more academic and less likely attacks.
- **Supply chain scanning** of components being in use under the agent, i.e., tools used to create prompts based on user initiative human-readable query, language used to create the chain, i.e., Python, and commonly used language do come up with underlying security vulnerabilities, so in short, we should consider the agent as a standalone box and do the full vulnerability scan.
- **Mitigations and Recommendations:**
  - Develop persona-based guidance (Developers, Security professionals, CISO, CxOs)for mitigating threats and securely adopting agentic systems.

  - Analyse the feasibility of human-in-the-loop interventions in critical systems and how to scale in multi-agent settings. Evaluate intelligent monitoring techniques and triadic adaptation, including anomaly detection and reinforcement learning from feedback (RLFH/RLAIF). Provide actionable recommendations for scaling secure agentic systems.
    - Include patterns. For example, AuthZ/AuthN patterns for different use cases—Authz/N per request/action? Actions taken as the user or as a system account? Point out where similarities exist with standard distributed system development vs. where the non-deterministic nature requires a different approach.
    -

- **Agentic Security Landscape:**
  - This is follow-up work to align with other work in this area
  - Map vulnerabilities and mitigation strategies to OWASP Top 10 for LLMs, OWASP AI Exchange, and Top 10 for Agentic Systems.
  - Cross-reference with MITRE, NIST, and industry guidelines
  - Regulatory compliance with legislation, including the AI Act.
  - Reference to Ethical and safety guidelines.
  - Highlight vendor solutions and tools for securing the agentic environment, if any.

Expected Outcomes & Artifacts

- **Agentic Security Threat Model and Vulnerabilities Taxonomy**
- **Agentic security sample implementations (exploits, intentionally vulnerable agentic apps)**
- **Securing Autonomous Agents Guide:**
  - Reference Model of Agentic AI Patterns
  - Reference to the previous two artefacts
  - Practical guidance for developers, security professionals, and policymakers.
- **Agentic AI Security Landscape Report**
  - Landscape of initiatives, tools, and vendors - this could be rolled into the main Solutions Guide Landscape.
  - Mapping and Integration with other AI Security Frameworks (OWASP, NIST, MITRE ATLAS, etc) - alignment strategies
- **Supporting references - for each other artifacts but also published and maintained centrally.**

Each outcome will have its own lead to help parallelise progress with ownership and cross-reference

Deliverable Roadmap

Deliverable	**Audience**	**Target Review, Publication Dates**
Agentic Security Threat Model and Vulnerabilities	Technologists, Developers, Security Researchers and other Security Professionals, Security	Review : January 2025 Publication: February 2025

	and related decision makers, CxOs, Chief AI Officer	
Securing Autonomous Agents Guide	Technologists, Developers, Security Researchers and other Security Professionals, Security and related decision makers, CxOs, Chief AI Officer	Review : January 2025 Publication : February 2025
Agentic security sample implementations (exploits, intentionally vulnerable agentic apps)	Technologists, Developers, Security Researchers and other Security Professionals	Review: April 2025 Publication: May 2025
A**gentic AI Security Landscape**	as above	Review : July 2025 Publication : August 2025
Agentic AI References	as above	Review : February 2025 Publication : March 2025 and ongoing

Contributing Team

- **Initial Core Contributors**
    - Andy Smith
    - Emile Delcourt
    - Emmanuel Guilherme
    - Evgeniy Kokuykin
    - Jason Ross
    - John Sotiropoulos
    - Helen Oakley
    - Krishna Sankar
    - Mohit Yadav
    - Patrik Natali
    - Rock Lambros
    - Ron F. Del Rosario
    - Sahana Chennabasappa
    - Sandy Dunn

    - Srinivas Inguva
    - Vin Giiarusso
    - Manish Kumar Yadav

- **Volunteer Request (if required):**
    - Developers, Vendors

Funding Asks, Requirements

- None, so far
- Salesforce AgentForce credits
- AWS Bedrock agent credits
- Twilio AI Assistants
- GPU Rental credits for open-source models

Notes from Core team voting:

References:

Vulnerable Autonomous Agents - OWASP Top 10 for LLM Applications - V2 Candidate Entry

Agentic Cookbook for Generative AI Agent Usage

Additional Team Literature Review

Draft Proposal for Top 10 for Agentic Systems
https://github.com/precize/OWASP-Agentic-AI

Vulnerable Autonomous Agent Threat Model -
https://github.com/jsotiro/ThreatModels/blob/main/LLM%20Threats-Autonomous%20Agents.png

LangChain - Autonomous Agents -
https://js.langchain.com/v0.1/docs/use_cases/autonomous_agents/

Imprompter: Tricking LLM Agents into Improper Tool Use - https://imprompter.ai/

Large Language Models On-Device with MediaPipe and TensorFlow Lite
https://developers.googleblog.com/en/large-language-models-on-device-with-mediapipe-and-tensorflow-lite/

On Device LLMs in Apple Devices: https://huggingface.co/blog/swift-coreml-llm

OWASP Top 10 for LLMs - Agentic Security Research Initiative

The AI Phones are coming
https://www.theverge.com/2024/1/16/24040562/samsung-unpacked-galaxy-ai-s24

Frontier AI: capabilities and risks – discussion paper -
https://www.gov.uk/government/publications/frontier-ai-capabilities-and-risks-discussion-paper

International Scientific Report on the Safety of Advanced AI -
https://www.gov.uk/government/publications/international-scientific-report-on-the-safety-**of-advanced-ai**

Here come the AI Worms - https://www.wired.com/story/here-come-the-ai-worms/

AI drone 'kills' human operator during 'simulation' - which US Air Force says didn't take place -
https://news.sky.com/story/ai-drone-kills-human-operator-during-simulation-which-us-air-force-says-didnt-take-place-12894929

Risks (and Benefits) of Generative AI and Large Language Models - Week 12 LLM Agents
- https://llmrisks.github.io/week12/

ENISA Report on Security and privacy considerations in autonomous agents -
https://www.enisa.europa.eu/publications/considerations-in-autonomous-agents

Integrating LLM and Reinforcement Learning for Cybersecurity-
https://arxiv.org/abs/2403.1767

Security and Efficiency of Personal LLM Agents - https://arxiv.org/abs/2402.04247v4

TrustAgent: Ensuring Safe and Trustworthy LLM-based Agents -
https://arxiv.org/abs/2402.11208v1

Prioritizing Safeguarding Over Autonomy: Risks of LLM Agents for Science -
https://arxiv.org/abs/2402.04247

AutoDefense: Multi-Agent LLM Defense against Jailbreak Attacks -
https://arxiv.org/abs/2402.11208v1

Workshop: Multi-Agent Security: Security as Key to AI Safety -
https://neurips.cc/virtual/2023/workshop/66520

Building a Zero Trust Security Model for Autonomous Systems -
https://spectrum.ieee.org/zero-trust-security-autonomous-systems

Securing LLM Backed Systems: Essential Authorization Practices
https://cloudsecurityalliance.org/artifacts/securing-llm-backed-systems-essential-authorization-practices

Adversarial Attacks on Multimodal Agents

OWASP Top 10 for LLMs - Agentic Security Research Initiative

https://chenwu.io/attack-agent/