

# OWASP Top 10 for Agentic Applications

## How Copilot Studio Helps Reduce Risk

The **OWASP Top 10 for Agentic Applications (2026)** identifies the most critical security risks introduced by autonomous, tool-using AI systems, based on community-driven analysis and real-world observations. Microsoft Copilot Studio integrates native guardrails and Microsoft Security services to help organizations mitigate these risks in real-world deployments.

### The OWASP Top 10 and Copilot Studio Mitigations

#### ASI01: Agent Goal Hijack

**Risk**  
Attackers inject hidden or manipulative instructions that redirect an agent's goals and multi-step behavior.

- Mitigations**
- Prompt filtering and default runtime guardrails block malicious instructions and goal deviations
  - Agents execute using the end user's Entra ID permissions by default, limiting blast radius
  - Microsoft Defender can help identify and block off-scope actions and exfiltration attempts

#### ASI02: Tool Misuse & Exploitation

**Risk**  
Agents misuse legitimate tools, causing data loss, excessive actions, or workflow hijacking.

- Mitigations**
- User-scoped tool access is selected by default, and makers are warned against the use of over-privileged service accounts
  - Administrators manage connectors/actions available to makers
  - Human approval can be built-in for sensitive actions
  - Full logging and alerting across tool usage with Purview and Sentinel

#### ASI03: Identity & Privilege Abuse

**Risk**  
Exploiting delegation, cached credentials, or role inheritance to escalate access.

- Mitigations**
- Agents typically operate under user identity and permissions, or within scopes of approved permissions
  - Warnings surface if elevated credentials are configured for use
  - Entra ID authentication, Conditional Access, and audit logging support least privilege and monitoring

#### ASI04: Agentic Supply Chain Vulnerabilities

**Risk**  
Malicious or tampered models, tools, connectors, or dynamically loaded components.

- Mitigations**
- Closed, admin-approved connector ecosystem
  - Closed, Microsoft-validated model availability
  - Connector allowlists and environment isolation
  - Sandboxed execution for actions and flows
  - Programmatic shutdown via Power Platform APIs

#### ASI05: Unexpected Code Execution

**Risk**  
Agents are manipulated into executing unintended or adversarial code.

- Mitigations**
- Custom logic runs only in managed, sandboxed services
  - Admin governance and runtime threat detection reduce injection paths
  - No arbitrary or OS level code execution by design

#### ASI06: Memory & Context Poisoning

**Risk**  
Persistent poisoning of stored context that affects future reasoning.

- Mitigations**
- Session memory does not persist
  - No persistent global memory
  - Agents use curated, protected, enterprise data sources governed by Purview and data loss prevention controls

#### ASI07: Insecure Inter-Agent Communication

**Risk**  
Weak authentication or integrity across agent-to-agent coordination.

- Mitigations**
- Single agent operation by default
  - Multi-agent connections require explicit configuration
  - All communication uses secured, authenticated APIs within environments that are isolated by default

#### ASI08: Cascading Failures

**Risk**  
A single error propagates across agents, tools, or workflows

- Mitigations**
- Limited autonomy and scoped execution by design
  - Monitoring, quotas, and alerts can act as circuit breakers
  - Full audit trails can enable rapid detection and containment

#### ASI09: Human-Agent Trust Exploitation

**Risk**  
Users over-trust confident agent recommendations and approve unsafe actions.

- Mitigations**
- Microsoft Purview sensitivity labels can be applied for sourced content, helping users judge trust and handling
  - Human-in-the-loop configurations can require users to review and confirm before executing actions
  - Automated security scans help flag risky configurations before publishing, encouraging secure best practices

#### ASI10: Rogue Agents

**Risk**  
Agents drift from intended purpose and behave like insider threats.

- Mitigations**
- Strong environment containment and data loss prevention enforcement
  - Comprehensive auditability that aids anomaly detection
  - One-click disablement and programmatic quarantine available
  - Agents cannot modify themselves or spawn new agents

Copilot Studio helps support a defense-in-depth approach, combining built-in identity, governance, runtime monitoring, and platform isolation measures to help organizations address many of the top risks highlighted by OWASP, helping to reduce the need for improvised controls and complementing existing security solutions.

[LEARN MORE](#)

Capabilities described rely on Microsoft Copilot Studio controls and Microsoft Security integrations; licensing and feature availability may vary.

OWASP Top 10 for Agentic Applications content © OWASP Foundation. This content is licensed under CC BY-SA 4.0. For more information, visit <https://creativecommons.org/licenses/by-sa/4.0/>