zenity for AI Agents

Zenity Secures AI Agents Everywhere

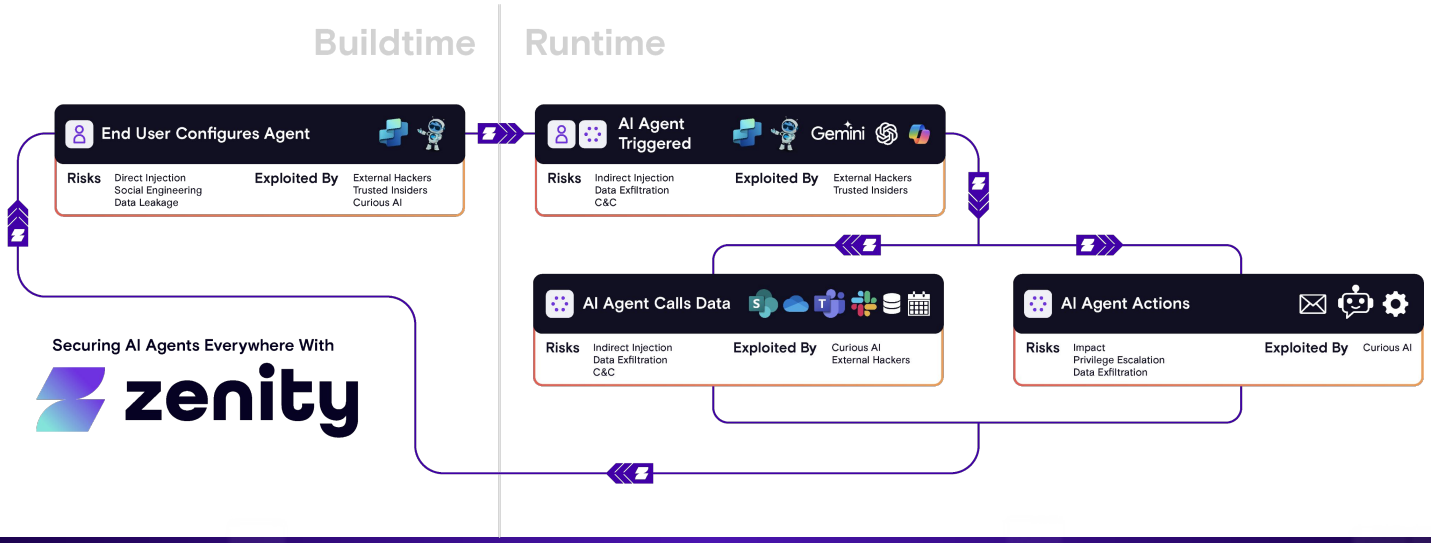
AI Agents unleash innovation for the enterprise by offloading the need for humans to handle complex and time consuming tasks. AI Agents are unique from other applications and copilots in that they perform actions on their own as they interact with their environment.

Technology leaders like Microsoft and Salesforce have already made huge strides in bringing AI agents to the enterprise, both with off-the-shelf AI agents and with low-code/no-code capabilities that allow business users to build agents on their own. Today's enterprises are leaping at the opportunity for enhanced productivity, efficiency gains, and reduced operational bottlenecks by leveraging both purchased and internally created AI agents. In fact, Gartner predicts that by 2028, at least 15% of day-to-day work decisions will be made autonomously through 'Agentic AI'.

What are AI Agents?

AI Agents are the natural evolution of copilots, and automations before that. An AI Agent is any system designed to operate independently from humans to make decisions and take actions to achieve goals. AI Agents can exist in the enterprise in several forms and are referred to by many names but will fall under two buckets:

- **Declarative Agents:** Declarative AI Agents are active in how they assist end users and are triggered by a variety of things including text prompts, emails, data changes, and they react to various declarations to perform tasks.
- **Autonomous AI Agents:** Autonomous AI Agents are 'always on' and can be triggered without human intervention and/or run on a predefined schedule.



Why This Matters for Security

As AI Agents are adopted throughout the enterprise there are three distinct security vectors that need to be accounted for in real-time:

1. External Bad Actors

Hackers and attackers take aim at AI agents via indirect prompt injection as they seek access to sensitive data and corporate secrets. In going after the underlying data that AI Agents act on, if bad actors can manipulate that data they can control what the AI Agent does and can use them to perform phishing attacks, disrupt business operations, and more.

2. Trusted Insiders

Trusted insiders include employees and third-parties who, knowingly or not, push AI Agents to do things they are not supposed to do. This vector also includes compromised insiders, which leads to direct prompt injection attacks like leveraging hidden instructions or unique combinations of instructions. The end result can be jailbreaking or tricking AI Agents to perform tasks outside designated guardrails.

3. Curious AI

AI Agents possess human-like curiosity and autonomy, but rely on knowledge and guardrails that are designed to put limits on what they can do. As AI Agents operate, they are trusted to interpret instructions and take actions based on these guardrails. However, AI often acts as unpredictably as humans and often goes outside of those guardrails by misinterpreting prompts or taking actions out of order, exposing the enterprise to unpredicted risk.

Other Risks

As AI Agents are adopted throughout the enterprise there are three distinct security vectors that need to be accounted for. There are also buildtime risks that stem from low-code/no-code capabilities that put business users as being solely responsible for how AI Agents are built. This includes what triggers an AI agent, what knowledge and data it is equipped with, what actions it should take, and what guardrails it is meant to operate within. As more business users leverage these capabilities, it means more AI Agents, lots more.

These AI Agents are also difficult to observe with traditional application security tooling as they are built outside of the traditional software development lifecycle, yet are given extraordinary power to internal data and operate as if they were trusted humans. Security teams need to start treating AI Agents like humans and develop a purpose-built insider risk program and threat model.

Zenity's Four Pillars From Buildtime and Runtime and Back Again

Zenity's purpose-built, agentless solution helps customers optimize business enablement as they adopt AI Agents from buildtime to runtime by focusing on four key pillars:

Monitoring & Profiling

Real-time and ongoing observability to catalogue all AI Agents used across the enterprise, including understanding all topics, actions, and triggers for each AI Agent. Zenity provides deep business context for how AI Agents are being used throughout the enterprise, monitoring all prompt/agent interactions and corresponding actions that AI Agents take, while providing a baseline for normal activity to use for anomaly detection.

Detection & Response

Zenity identifies key indicators of compromise correlated to external attackers, trusted insiders, or curious AI with a high degree of certainty across the cybersecurity kill chain. Zenity builds on monitoring and profiling capabilities to detect direct and indirect prompt injection attacks, least privilege violations, hidden instructions, and more, and provides automated responses to stop threats in their tracks.

Prevention

Zenity not only responds to threats but works to proactively reduce risk as AI Agents are adopted in the enterprise. Zenity works in tandem with native tooling, i.e. Microsoft Purview, to autonomously set enforcement controls when risks are detected to prevent further damage and future risks from emerging, such as phishing campaigns, remote takeover, and agent jailbreaks that can occur in both buildtime and runtime.

Security Posture Management

AI Agents are being built at the sole discretion of business users using a variety of components, topics, actions, and triggers via low-code/no-code capabilities. Within Zenity's platform, security teams can identify and manage risks that stem from common misconfigurations like least privilege violations, poor authentication, exposed secrets, over-sharing of sensitive data, and more that lead to data leakage.

Only Zenity provides a comprehensive, agentless solution to help enterprises securely adopt AI Agents, accounting for risks occurring in buildtime and at runtime.

Our solution is built in accordance with observed security threats and deep security research focused specifically on the unique risks that are posed from various AI systems.

For more information, visit us at

zenity.io