NetSPI™ + Microsoft

How Microsoft and NetSPI Partnered to Build a Standardized AI Security Framework Securing 70+ Products

Microsoft is trusted by millions worldwide to deliver innovative cloud, productivity, and AI solutions that empower people and communities to achieve more.

The Challenge

NetSPI Solutions
AI/ML Security Assessment Framework

Industry
Technology & Software / Cloud & AI

Employee Count
220,000-250,000

Headquarters
Redmond, Washington

Website
microsoft.com

As a global leader in artificial intelligence, Microsoft's highest priority is delivering AI models and solutions that customers can trust. This requires a deep understanding of the risks introduced with AI-based solutions. Microsoft developed a comprehensive categorization of AI vulnerabilities, published by the **Microsoft Security Response Center (MSRC)**, to define the types of threats and their potential impact.

However, creating a standard is only the first step. Microsoft's Principal Security Assurance Engineer, Daniel Moore, needed a way to validate the resilience of its AI implementations against the new class of vulnerabilities **introduced by AI**. The company performs security assessments through various channels, including internal teams and third-party specialists. Yet, for AI, there was no established standard to measure against. Without a structured framework, any testing would be ad-hoc, making it impossible to establish scope, gauge severity, or ensure quality coverage. Beyond addressing technical vulnerabilities, Microsoft needed a framework that ensured its AI solutions were not only secure but also ethical, trustworthy, and safe for public use. This methodology had to account for risks that went beyond the purely technical—ensuring AI systems would not inadvertently enable harmful behavior or produce inappropriate outputs. This lack of a safe repeatable methodology meant Microsoft could not be fully confident in the security posture of its AI solutions or effectively communicate that resilience to customers.

"In the absence of structured criteria, there is no ability to establish scope, severity, or even quality of coverage. Any test effort for AI vulnerabilities would be completely ad-hoc and would in no way give us confidence in the posture of our AI implementation."

Daniel Moore, Principal Security Assurance Engineer, Microsoft

NetSPI™ + Microsoft

The Solution

Microsoft required a partner with specialized AI security expertise to co-develop a testing framework that could verify the resilience of its AI implementations. Having an established multi-year relationship with NetSPI that had already produced verifiable results, Microsoft selected them for this critical initiative. NetSPI brought strong AI expertise to the table and demonstrated the flexibility and willingness to collaborate on building the AI/ML Security Assessment Framework from the ground up.

The development was a multi-phased, collaborative process:

Establishing Goals: The teams first aligned on the available expertise and the desired outcomes for the framework.

Leveraging Existing Work: They utilized existing work products to reinforce thinking around scope, coverage, and assessment criteria.

Expert Review: The draft framework was rigorously reviewed by Microsoft's internal AI experts.

Testing and Tuning: Finally, the framework was tested in real-world scenarios and tuned based on the outcomes to ensure its effectiveness.

This joint effort produced a structured, repeatable, and measurable standard for assessing AI security. The AI Testing Framework provides comprehensive coverage when performing AI assessments, enabling Microsoft and its customers to gain confidence in the security and resilience of its AI solutions.

The Results

The implementation of the AI/ML Security Assessment Framework delivered significant value, providing a consistent and measurable method for evaluating Microsoft's AI security posture.

- **Established and Repeatable Testing:** The framework has been highly effective, creating a structured process with established criteria for AI security testing. This allows Microsoft to evaluate its overall AI posture consistently and repeatably. As Daniel Moore at Microsoft noted, ***"In the absence of structured criteria... any test effort for AI vulnerabilities would be completely ad-hoc and would in no way give us confidence in the posture of our AI implementation."***
- **Extensive Vulnerability Identification:** The framework has been integrated into Microsoft's ongoing penetration testing program. NetSPI has performed 28 tests for Microsoft using the framework, informing over 70 products. This structured approach ensures thorough coverage and allows for the methodical identification and scoring of vulnerabilities. Through these tests, NetSPI discovered 126 vulnerabilities (Critical, High, Medium, etc.), which then follow the same rigorous mitigation and retest protocol used for all security findings.

- **Enhanced Customer and Industry Trust:** With a standardized framework, Microsoft can provide clear evidence of completed testing, outcomes, and mitigations to existing customers and new opportunities. This demonstrable commitment to security builds confidence and reinforces Microsoft's position as a leader in trustworthy AI.
- **Proven Partnership and Expertise:** The collaboration highlighted NetSPI's ability to adapt to emerging business requirements and invest in delivering needed outcomes. ***"There is a steep learning curve associated with gaining competence in this space,"*** Moore explained. ***"NetSPI has gained this expertise with the scrutiny and collaboration of Microsoft. You want someone doing this work in whom you have confidence is capable of delivering quality results."***

For organizations looking to advance their AI/ML security, Moore advises against reinventing the wheel. ***"Look for trusted and established approaches for assessing your AI/ML security."*** The partnership between Microsoft and NetSPI showcases the power of combining deep industry knowledge with specialized security expertise to build a framework that not only solves a critical business problem but also sets a new standard for the industry.

"We have a consistent, measurable, and repeatable method of assessing and communicating outcomes for our AI testing."

Daniel Moore
Principal Security Assurance Engineer
Microsoft

"To date, we have performed the AI assessment as an integrated part of our ongoing pen testing. This has been completed for about 70 product tests over the last two years."

"NetSPI has demonstrated the ability to listen and adapt as needed to emerging business requirements. They have consistently invested in ways that ensure their effectiveness in delivering the outcomes we need."

Contact us to advance your AI/ML security

www.netspi.com/contact