S&P Global
Market Intelligence

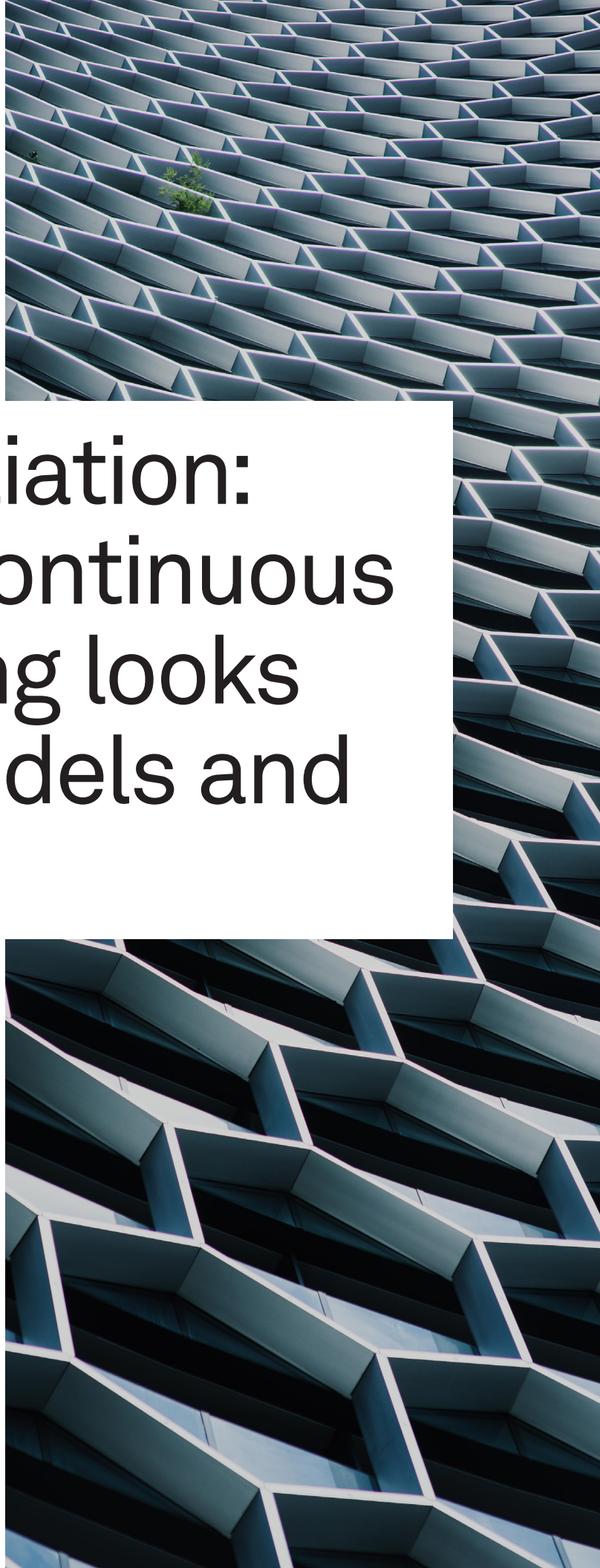
451 Research Market
Insight Report Reprint

Coverage Initiation: Mindgard's continuous AI red teaming looks to secure models and applications

April 30, 2025

by Justin Lam

Critical to modern app development, incorporating the AI model layer requires understanding and mitigating the weaknesses and risks for underlying services and models. The combination of new risks and new opportunities for vendors and their enterprise customers will require additional approaches.



Introduction

While the discipline of red teaming is not new, with security teams actively emulating adversarial tactics, techniques and procedures, Mindgard looks to primarily use a technology and automation-first approach with humans in the loop for additional reinforcement, as the verification of generative AI harms like large language model jailbreaking, bias or abuse requires human interpretation for now. The company is providing red teaming for generative AI rather than simply automating GenAI tools for conventional red teaming or penetration testing.

THE TAKE

Continuous testing of GenAI in models and applications will remain critical. Newer GenAI-based services may not need to cross the same competitive moats as previous generations of cloud and SaaS, so Mindgard's addressable market surface area could grow geometrically. Although recent innovations, such as Anthropic's Model Context Protocol and Google's Agent2Agent, may streamline interfaces, standardize security between agents and accelerate integration between multiple models, the potential growth for attack surface area from poorly implemented integrations could be very high. The underlying GenAI red teaming needs to be continuous in part because the entire ecosystem is changing at all levels. Profound continuous changes in unit, user and technology provider economics; technology; the harms; and adversaries will require constant adjustment. Mindgard must continuously sharpen and adjust its focus to add immediate value and refine the iterative approach needed for success.

Context

Recent 451 Research reports have noted the transition from "software as a service" to "service as software" that promises to disrupt market economics for incumbent vendors and their enterprise customers. Core to this disruption is the use of agentic interfaces that actually perform knowledge work in addition to better supporting knowledge workers. The potential disruptions look to improve collective productivity between users and their systems.

Company

Mindgard was founded in May 2022 by Peter Garraghan of Lancaster University and entrepreneur Steve Street. Garraghan is a professor of computer science, focusing on distributed systems and AI and is winner of the EPSRC Fellowship, a UK-based grant that recognizes and cultivates further research into AI/machine-learning security. Based in London, Mindgard has raised a total of $12 million in seed and initial venture investment since September 2023 from investors including IQ Capital Partners, .406 Ventures, Atlantic Bridge, Lakestar, Osney Capital and WillowTree Investments. With about 15 employees, the company's personnel are heavily focused on research, development and implementation.

Services/Technology

Mindgard's focus on a productized service differs from conventional red teams that are often led by human penetration testers. While pen testers use many automated tools to simulate adversaries and find weaknesses, Mindgard wishes to invert priorities, leading with a technology and automation-first approach, delivered as a service.

For customers, the first step in their Mindgard journey requires pointing the Mindgard platform to existing AI models and applications. A suite of automated tests checks for several potential GenAI security and safety issues such as model extraction, prompt injection, inversion, evasion and poisoning. Blended attacks against data confidentiality, integrity and availability are simulated to observe or detect data leakage, intellectual property theft or outright large language model jailbreaking.

From there, detailed views on the scenarios executed and attack paths are collected and reported via existing workflows in security information and event management, ticketing or reporting tools, and reports are enriched to OWASP and MITRE standards as well as with suggested mitigations. Analogous to more traditional red teaming approaches, Mindgard leads with technology first to reduce dependency on any individuals or personnel. Historical red team and pen testing assignments have been defined by billable hours, and Mindgard's continuous red teaming seeks to reduce that dependency.

Wisely, Mindgard still maintains a human in the loop. Unlike other red teaming approaches, potential issues to data integrity, trust and safety require human interpretation. Issues with model bias, abuse or safety cannot be deterministically interpreted or evaluated by technology alone.

Strategy

Mindgard's strategy to be a technology-first AI red teaming product enables interesting go-to-market choices. While initial ideal customers have self-identified as those with significant GenAI investment and growing expertise, many early adopters may have already tried other homegrown or community-sourced approaches and look to augment their efforts with a more automated approach to implement and deliver their own applications and models faster.

For Mindgard, the underlying philosophical value driver is a customer's time to market. Whether it is an independent software vendor or technology company that looks to provide a GenAI-based service or a large enterprise looking to internally adopt GenAI applications to boost organizational productivity faster, reducing customer time to market is a foundational value in Mindgard's mission.

Other interesting opportunities could include distribution with personnel-led quasi-competitors. Bug bounty program providers such as HackerOne, Synack, Cobalt and Bugcrowd have marshaled and vetted many independent security researchers, and while these researchers and red teamers have brought their own tools like Burpsuite, semgrep or NMAP, it is possible that they could implement or leverage Mindgard in their own engagements.

The research required to keep up with potential GenAI security issues is also significant, and the importance of Mindgard's connection to research organizations like Lancaster University and EPSRC could be a major competitive advantage. Like 2024's RSAC Innovation Sandbox winner Reality Defender with its oversized research team, there will be heavy lifting required to keep up with potential GenAI risks. GenAI fire will be required to fight GenAI fire; research at the forefront is needed to stay ahead of the cat-and-mouse contests with adversaries.

The shifting unit, user and technology economics that GenAI could bring about could also create interesting partnerships in the future. Emerging GenAI services that wish to show how their services can be used and operated safely may employ Mindgard services to vet or refine their choices. Changing technology interfaces, such as Anthropic's Model Context Protocol or Google's Agent2Agent, further disrupt how Mindgard partners or inserts itself into the market. Hugging Face lists 1.5 million language models for integration. All these changes are opportunities for Mindgard and they do bolster the case for continuous red teaming.

Competition

As a discipline, red teaming has evolved from pen testing and periodic assessments. A generation ago, much pen testing was driven by compliance mandates like PCI-DSS or HIPAA. Today, more severe and sophisticated attacks involving lateral movement and lengthier "kill chains" require additional approaches to simulate rapidly advancing tactics, techniques and procedures, driving merging of red and blue teams, using their knowledge of TTPs to simulate or detect adversaries.

Existing security service companies with large teams of experts, assessors and auditors are beginning to offer AI red teaming services. Aforementioned bug bounty platforms are also offering rewards for bugs or vulnerabilities found in GenAI systems.

Homegrown tools and open-source tooling such as PyRIT (Python Risk Identification Tool) are indirect competitors in that they require a greater level of integration effort, resulting in extended implementation time and increased maintenance burden over a commercially delivered and supported service. Deep AI red teams at major platform providers such as Microsoft Corp. are not necessarily a direct competitor, but they are a broader part of Microsoft's Secure Future Initiative.

Security for GenAI has attracted a number of new vendors, including Hidden Layer, Witness.ai, Robust Intelligence, Protect.ai, Calypso.ai and many more. Other data security vendors ranging from Varonis, Cyera, BigID and Proofpoint are leveraging their offerings in data classification and discovery to enable safer GenAI adoption. Security of GenAI does not have neatly defined segments compared with more mature security segments; the fight for wallet share is inevitable. In 451 Research's Voice of the Enterprise: Information Security, Technology Roadmap 2024 survey, security for generative AI was the most-cited technology in pilot/proof of concept or plan to deploy in the next 6–24 months.

SWOT Analysis

STRENGTHS

Mindgard's focus on red teaming with a technology-first approach could enable larger distribution among existing red teams and security service providers. The company's strong connection to research organizations reflects the need for cutting-edge development, given the shifting landscape.

WEAKNESSES

Identifying the ideal customer profile demographic is just the first step. Success with Mindgard is more about the process of adoption than the adopted product, and it is imperative that Mindgard master both the customer buying journey and the customer success journey over time.

OPPORTUNITIES

The entire agentic economy promises to disrupt conventional SaaS economics as software as a service is replaced by service as software. Harnessing the need for faster and more iterative time to market in this ecosystem could be a massive opportunity.

THREATS

The number of "security for GenAI" companies is skyrocketing. While Hidden Layer was the only such entrant in the 2023 RSAC Innovation Sandbox, this year's competition features no fewer than eight "security for GenAI" finalists.

CONTACTS

Americas: +1 800 447 2273
Japan: +81 3 6262 1887
Asia-Pacific: +60 4 291 3600
Europe, Middle East, Africa: +44 (0) 134 432 8300

www.spglobal.com/marketintelligence
www.spglobal.com/en/enterprise/about/contact-us.html