Cool Vendors for AI Security

23 October 2024- ID G00818565- 12 min read

By Jeremy D'Hoinne, Bart Willemsen, **and 2 more**

Security and risk management leaders face operational challenges, compliance concerns and attacks aimed at AI systems. This Cool Vendor research identifies providers with innovative ways of securing AI applications, supporting AI trust, risk and security management capabilities.

Overview

Key Findings

- Security and risk management leaders need to upgrade data and application security practices to prevent incidents and potential breaches targeting new AI applications that their organization develops.
- Frameworks and guidelines are available, such as the AI trust, risk and security management (AI TRiSM), which highlights the functional areas for mitigating AI and cybersecurity risks. However, AI security providers offer overlapping and immature capabilities, adding complexity to the identification of the necessary controls.
- Application and data security teams often conduct evaluations in silos, leading to data protection tools being purchased as stand-alone products. This lack of coordination often means that AI security testing technologies are lacking to confirm effective data security practices.
- Technical advisors in security teams lack benchmarks, peer feedback or methodology to evaluate what is good enough security. This is detrimental to good evaluation, as product demos can be impressive but do not necessarily translate into an efficient deployment at scale.

Recommendations

Security and risk management leaders involved in securing enterprise AI initiatives should:

- Build sufficient AI literacy through training and knowledge sharing in the security teams. Ensure that teams involved with AI applications at design, development and runtime can identify new attack surfaces and gaps in existing defense.

- Conduct pilots and experiments to evaluate how dedicated AI runtime defense technologies help prevent AI abuses leveraging prompt injections, or attempt to jailbreak AI applications beyond their original scope.
- Mitigate privacy and data security issues by integrating data protection techniques early in development, and AI security testing in the continuous integration/continuous delivery (CI/CD) pipeline.
- Initiate improvements at each step without neglecting security operation teams. Managing new alerts related to AI systems will require monitoring changes.

Strategic Planning Assumption

Through 2025, generative AI will cause a spike of cybersecurity resources required to secure it, causing more than a 15% incremental spend on application and data security.

Analysis

This research does not constitute an exhaustive list of vendors in any given technology area, but rather is designed to highlight interesting, new and innovative vendors, products and services. Gartner disclaims all warranties, express or implied, with respect to this research, including any warranties of merchantability or fitness for a particular purpose.

What You Need to Know

In addition to conventional risks on data and application security, all of which remain relevant, the use of AI technology inherently brings more specific risks with it. Enterprises increase their efforts to consume and build AI applications. They need to implement AI guardrails and specific privacy, quality, reliability and security controls across all phases of the technology life cycle. AI TRiSM supports this multifaceted challenge for security and risk management leaders who are involved in supporting their organization’s AI initiatives, and their needs to collaborate with AI leads to implement required security and safety measures.

Gartner observes technology innovations in key areas of AI TRiSM across the secure software development life cycle (SSDLC) with a portfolio of products, frequently including:

- AI runtime defense (AIRD): These tools aim at providing the “AI runtime enforcement” capabilities of AI TRiSM for enterprise-developed applications. Today, they primarily provide near-real-time security controls against new forms of attacks targeting AI applications, such as prompt injections and jailbreaks. They often include more minimal AI content anomaly detection (“guardrails”) to detect harmful and toxic inputs, or even hallucinations and profanity. Some providers offer specialized versions of these tools (e.g., AIRD for M365 Copilot, AIRD for GitHub Copilot).
- AI security testing (AIST): These technologies automate the creation and operations of automated testing of inputs and outputs for AI systems. A growing number of technologies specialize in supporting the security team’s effort when they perform AI red teaming exercises for large language model (LLM) chatbots. They offer tools to automatically test the resistance of the model against adversarial inputs, but also to verify the impact of controls implemented in the AI application stack or in front of the AI chatbots.

Other capabilities are available, such as specialized controls to monitor and enforce policies based on employees' activity consuming commercial third-party AI applications (e.g., ChatGPT), or model management tools focused on alignment with security standards, such as NIST AI RMF. This Cool Vendor document can only highlight a handful of providers delivering these capabilities, but there is no shortage of coolness in the AI security market space.

Security and risk management leaders should expect feature overlaps and blurred boundaries between tools. AI TRiSM gives a broader overview of all the capabilities required to effectively govern AI trust, risk and security. It is important to ensure that application security teams operate under the guidance of AI cybersecurity governance principles.

With such a massive influx of innovations, Gartner released the Innovation Guide for Generative AI in Trust Risk and Security Management and Cool Vendors in Data Security, including many other interesting companies and promising features. Similarly, despite ongoing progress and available offerings from network and cloud security providers, this document focuses on specialized AI TRiSM providers protecting enterprise applications. This document also doesn't cover uses of AI for cybersecurity (see How to Evaluate Cybersecurity AI Assistants for more information).

HiddenLayer

Austin, TX, U.S. (hiddenlayer.com)

Analysis by: Jeremy D'Hoinne, Avivah Litan

Why Cool: HiddenLayer roots of coolness lie in its approach to testing AI security across the life cycle of AI applications. Its suite of products include adversarial model testing, which got recognition earlier this year following the integration of their model scanner in the Microsoft Azure AI Studio catalog.[1] HiddenLayer also provides security posture management controls, scanning model artifacts, checking model integrity and mapping the results of the scans with Mitre Atlas TTPs.

By offering capabilities not only for LLM applications, but also other types of AI assets, HiddenLayer can test multiple types of models in areas such as computer vision or credit fraud. The vendor also offers to scan public model repositories to detect malicious or compromised models.

HiddenLayer also offers runtime AI defense, not limited to generative AI (GenAI) applications, and a newer prompt monitoring component. It can detect and report anomalies in real time, including attacks such as prompt injection or jailbreaks, personally identifiable information (PII) leakage and also types of attacks intending to alter the model behaviors. The vendor is in the process of releasing an AI security testing module as part of its platform, with availability expected in 2024.

Challenges: As a startup trying to address multiple areas of AI security and going beyond LLM-only products, HiddenLayer must maintain a strong development pace while keeping an edge on more traditional security vendors. The vendor needs to streamline its user workflows across the various dashboard and views to facilitate collaboration between the multiple roles that could leverage its solution.

Who Should Care: Chief information security officers (CISOs) who must protect AI applications at runtime and need to ensure that AI teams adequately implement security testing during the entire AI development life cycle.

Holistic AI

London, U.K. (holisticai.com)

Analysis by: Bart Willemsen, Avivah Litan

Why Cool: Holistic AI's cool factor lies in the detailed and meticulous process to monitor, control and reduce risk in AI applications at multiple stages of AI application development. The vendor provides security and risk management features, including bias detection, robustness and privacy risk assessments, for third-party or first-party models, and supports intervention to enhance model accuracy and fairness. Holistic AI supports specialized AI model audits for various regulations, such as the EU AI Act and NYC bias detection regulation.

Holistic AI supports privacy protection by generating and using synthetic data to test and audit algorithms. The algorithm auditing capabilities are flanked with a governance platform (AI discovery, inventory, monitoring and risk management), compliance overviews and regulation tracker for global and regional insights. Holistic's auditing controls help detect issues on privacy, explainability and fairness, profanity and toxic language, in proprietary as well as third-party provisioned GenAI models.

Finally, Holistic AI offers global regulatory insights through a comprehensive knowledge base and promotes awareness through AI governance training activities. Practical guidance includes sector (e.g., state- or country-specific risk mitigation information).

Challenges: With many vendors emphasizing the use of AI in augmentation of functions, it stands out that there is no granular chatbotlike interface on Holistic AI's "tracker" module, which would make getting the right information sought after from a global covering repository much more accessible. This will only get worse since it's expected the amount of needed resources and legal interpretation or references will grow with great speed for the foreseeable future.

Who Should Care: Leaders in security and risk management or data and analytics responsible for AI governance, trust, compliance and safety should evaluate Holistic AI. They must implement AI governance technology to support enterprise policies and regulatory compliance.

Lasso Security

Tel Aviv, Israel (lasso.security)

Analysis by: Dennis Xu, Jeremy D'Hoinne

Why Cool: Lasso Security offers specialized LLM security and safety features with a focus on protecting against employees misuse, error or malicious activity. This includes whether they consume third-party chatbots, organizations developing LLM-powered applications or developers using IDE with embedded GenAI features like GitHub Copilot.

The vendor's dashboards enable discovery of most consumed apps and users with higher-risk behaviors. Lasso's management console supports custom topic moderation based on user role and query and response contexts expressed in natural language, with in-line content redaction and predefined categories, such as detection of sensitive information. Its specialized GitHub controls can highlight and mask sensitive data shared with the code assistant's servers, and provide point-in-time security scans.

Challenges: Lasso's chatbot protection product primarily deploys as a browser extension, which might create friction in some organizations. The vendor is small, and has recently expanded toward protecting custom enterprise AI applications. Choosing between stronger focus and expanded coverage is one of the most difficult decisions for a startup, and GenAI startups like Lasso will need to make the right choice — perhaps multiple times in the next few years.

Who Should Care: Security and risk management leaders involved in enforcing GenAI usage policies need to understand and investigate the benefits of a solution going beyond application basic categorization to better understand what users really do. Application security leaders should investigate the controls focused on custom enterprise AI applications.

Privya

Tel Aviv, Israel (privya.ai)

Analysis by: Bart Willemsen

Why Cool: Privya automates privacy assessments in software development for early detection and resolution of data protection violations. Its solution scans code nonintrusively to detect what (personal) data is used, how, by whom and thus detecting privacy violations at the earliest moment in operation.

Its LLM usage enables ingestion of code as natural language, verification of and matching against purpose lists to assess correct use of data, and creation of a risk overview and remediation actions. The training of its model includes compliance controls from well over 100 global and regional standards, including NIST and EU AI Act standards. Having been trained on several open source and Github repositories, the resulting vectors allow the LLM to detect similarities across operations with single scans.

Integrated in the CI/CD pipeline, the scanning is done from code repositories and thus without impact in production. At every commit, code is pulled, scanned, analyzed for data usage and, after potential violations are detected, the code is deleted without content retention to ensure its confidentiality. Violations can include improper use of (personal) data, credential compromise or unintended inclusion of any third party or AI model inclusion.

This type of "shift-left" approach is part of privacy engineering and advances privacy controls to the application or process development and operation phase. Barriers to shifting left usually have been capacity (accessing the code without obstructing operations) and cost (investing in the review of potentially thousands of lines of code and making surgical changes).

Challenges: As is the case with GenAI these days, accuracy is of concern with every LLM inclusion. Applicability for third-party provisioned, democratized (Gen)AI is clear to see, but the feasibility on noncontrolled models remains ambiguous at best. The user has to have access to the code, which will thus not work with third-party-provisioned GenAI, for example, which often remains a black box.

Who Should Care: Application and security and risk management leaders, in charge of ensuring compliance and assessing privacy risk within the applications they develop, should evaluate Privya's capabilities. Many painstaking efforts have been made to retrospectively make sense of data processing activities, yet proactive controls through shift-left privacy assessments of code must be on everyone's visor.

This includes detecting risk of unauthorized use of third-party technology, untrusted AI model usage and confidentiality risks on nonpersonal data (e.g., passwords). An added feed to data discovery efforts and records of processing activities are bonuses for those wishing to further automate privacy capabilities.

Robust Intelligence

San Francisco, CA, U.S. (robustintelligence.com)

Analysis by: Jeremy D'Hoinne

Why Cool: Robust Intelligence leverages offensive security research, threat intelligence and machine learning to algorithmically create tests and controls for enterprise AI applications. This combination helped Robust Intelligence create an adversarial model testing product ("AI validation"), deployed at the development and preproduction steps.

The vendor's involvement in AI security standard creation also shows in the runtime detection product, where AI safety (e.g., toxicity) and security (e.g., prompt injection) detections list OWASP top 10, National Institute of Standards and Technology (NIST) and Mitre Atlas categories directly in the various dashboards. Robust Intelligence's deployment options enable it to perform controls and verifications that include the RAG pipeline, or check consistency against the enterprise vector database.

Challenges: Robust Intelligence does not offer specialized tools to protect out-of-the-box AI application consumptions like ChatGPT (it can do it with API integration) or embedded enterprise AI applications like Copilot. In August 2024, Cisco announced its intent to acquire Robust Intelligence, and completed the acquisition right before publication of this document. This might influence future roadmap and Robust Intelligence prospects decisions.

Who Should Care: Security and risk management leaders, in charge of implementing secure development and runtime controls for enterprise AI applications, should evaluate how to work with their AI counterparts to add adversarial testing. Implementing runtime controls will require additional efforts to build the monitoring and incident response process. Deploying detection capabilities early on, with a vendor like Robust Intelligence, could help eliminate the most common issues automatically.

Evidence

[1] HiddenLayer Collaborates With Microsoft Azure AI to Enhance Model Security, PR Newswire