No.86
May 2025

HardenStance Briefing

Trusted research, analysis & insight in IT & telecom security **PUBLIC/UNSPONSORED**

AI Highlights from RSAC 2025

HardenStance attended RSAC 2025 last week. AI dominated once again, as it did last year. The emphasis this year was on the safeguarding of employees' use of Gen AI applications as well as introducing Agentic AI into cybersecurity operations.

- Palo Alto Networks and Cisco made announcements on protecting organizations against both developers and users exposing them to risk with their use of Gen AI. Akamai announced a new 'Firewall for AI'. Qualys announced new features for enabling more secure use of AI in development environments. Sandbox AQ's next release will also have as yet undisclosed features for protecting AI deployments.
- Google and CrowdStrike announced AI agents with potential to proactively and autonomously ask, answer and act on relevant questions in cybersecurity operations. These can unleash the power of AI way beyond Gen AI merely responding to a human SOC analyst's question in natural language. Palo Alto Networks' Unit 42 published research on the detailed nature of the threats posed by AI agents and how to defend against them.

Here are some of the highlights of RSAC 2025 in terms of new product announcements for protecting organizations against AI risk and using AI for cybersecurity use cases:

Akamai

Akamai announced a 'Firewall for AI' providing multilayered protection for AI applications against unauthorized queries, adversarial inputs, and large-scale data-scraping attempts. The company is pitching it as a purpose-built security solution designed to protect AI-powered applications, LLMs, and AI-driven APIs from emerging threats. By securing inbound AI queries and outbound AI responses, the firewall closes security gaps that generative AI technologies introduce.

Cisco

Cisco announced that "AI Defense", announced in January, is GA now. AI Defense is designed to be able to understand and supervise AI so that organizations can safeguard against the misuse of AI apps by their developers and users. It also protects against increasingly sophisticated threats to organizations from AI systems. Cisco reports seeing malware embedded in a number of popular Gen AI apps that its customers are routinely relying on. AI Defense is designed to be leveraged as a feature in an organization's existing Security Service Edge (SSE) or firewall platforms, hence also in any switch or router to allow AI Defense to be embedded in a customer's network. Cisco also announced Service Now as a go to market channel partner for AI Defense.

CrowdStrike

CrowdStrike unveiled 'Charlotte AI' Agentic Response and 'Charlotte AI' Agentic Workflows as features on its new Charlotte AI platform. Charlotte AI Agentic Response automatically asks and answers the investigative questions a security analyst would pose, accelerating root cause analysis, mapping lateral movement and guiding next steps. 'Charlotte AI' Agentic Workflows are drag-and-drop, LLM-powered workflows that enable analysts to insert and activate AI reasoning directly into automated playbooks.

Google

Google announced a number of new AI agents for various components of Google Unified Security. These are a natural language parser extension; a Detection Engineering Agent for automated rule creation and testing; a Response Agent to generate automation playbooks; an Alert Triage Agent in Google Security Operations and a Malware analysis agent in Google Threat Intelligence. Some are already available, others are to be released later this year.

Palo Alto Networks

Four AI-related product enhancements were among the most eye-catching in a flurry of RSAC announcements by Palo Alto Networks:

- **The launch of a new AI security platform - Prisma AIRS.** This is designed to protect an organization's entire enterprise AI ecosystem – AI apps, agents, models, and data. Key features include AI model scanning, security posture management of an organization's AI ecosystem; red teaming; run time security and AI agent security. The company also announced the intent to augment Prisma AIRS via the acquisition of Seattle-based Protect AI.
- **Reduction in organizational risk posed by misuse of Gen AI-driven apps with a new release of Prisma Access Browser**. This enhancement enables the safe use of Gen AI-driven applications in organizations by protecting against their misuse by users. Whether it's a coding assistant, writing assistant or other Gen AI app, the new release of Prisma Access Browser provides a variety of different controls for categorizing and applying a variety of policies to more than 2,000 Gen AI apps. Examples include blocking usage of an unapproved Gen AI app outright or redirecting users to one that's approved.

  This is just one feature in the differentiated positioning of a natively integrated Browser within Prisma Access and Prisma SASE. This leverages the browser's ability to see into traffic where network based controls increasingly can't due to the growth in application encryption.

The new release of Prisma Access Browser provides a variety of different controls for categorizing and applying a variety of policies to more than 2,000 Gen AI apps.

- **Gen AI-driven content analysis of emails within Cortex XSIAM 3.0.** News that the development team working on XSIAM, Palo Alto Networks' security operations product, has extended its coverage to include email security isn't surprising. To counter even the very best written phishing emails that threat actors can generate using Gen AI, X-SIAM now uses Gen AI in turn to inspect and accord emails a risk score based on the intent it can infer from the language in the email. This takes into account factors like the urgency with which a response is being sought and any requirement to trigger a financial transaction.

  Combined with additional file and URL context, including from more than 70,000 customers of the company's Wildfire assets, this inference of intent augments the known virus, known file, link scanning and spam controls that are already baked in to most cloud email solutions.
- **New research on the threat posed by AI agents**: On the last day of RSAC, Palo Alto Networks' Unit 42 threat research group published research and recommendations on how to manage the risk posed by AI agents, whether they're used for cybersecurity or other use cases. The research presents nine attack scenarios that can result in outcomes such as information leakage, credential theft, tool exploitation and remote code execution. It shows that most vulnerabilities and attack vectors are largely framework-agnostic, arising from insecure design patterns, misconfigurations and unsafe tool integrations, rather than flaws in the frameworks themselves. The research also presents mitigations for each one. Unit 42 has open sourced the source code and data sets on GitHub.

Qualys

Qualys focused a lot of its RSA presence on 'Total AI', an LLM scanner integrated into its Web Application Scanner (WAS) infrastructure as announced last December. With the latest updates, Total AI now provides comprehensive security testing of LLMs that are hosted on-premises with internal access only — i.e. without exposing models externally; detection of more than 38 jailbreak and prompt manipulation attack scenarios that can alter LLM behaviour; AI supply chain protection – continuous monitoring to detect hallucination attacks where LLMs are tricked into recommending non-existent but malicious third party packages; multimodal threat detection that identifies prompts hidden inside images, audio, and video files that are designed to manipulate LLM output; and Universal endpoint scanning - one-click security assessments for any LLM exposing an OpenAI-compatible chat-completion API, including AWS Bedrock, Azure AI, Hugging Face, Google Vertex AI, and self-hosted deployments.

Sandbox AQ

Until the week before RSAC 2025, Sandbox AQ's AQtive Guard solution mainly leveraged its foundational Large Quantitative Model (LQM) software for cryptography management. With a major new release, now available as SaaS, Sandbox AQ is seeking to disrupt the identity management space too. This new release leverages cryptographic identities as well as other types of artifacts to extend monitoring and management to diverse human and non human identities. AQtive Guard can now automate remediation as well as discovery of vulnerabilities and threats across both these key domains.

The strategy goes beyond just a combined crypto management and identity management play. It's a broader AI SecOps play, complete with integrations with Palo Alto Networks and CrowdStrike. The next release targets monitoring and managing how secure an organization's AI posture is (or isn't). ■

More Information

- Video interview: "A New Release of Sandbox AQ's AQtive Guard" (February 2025)
- Briefing: "New Standards for Securing AI" (November 2020)
- White Paper "AI in Cybersecurity: Filtering out the Noise" (February 2019)

About HardenStance

HardenStance provides trusted research, analysis and insight in IT and telecom security. HardenStance is a leader in custom cyber security research and leading publisher of cyber security reports. HardenStance is also a strong advocate of industry collaboration in cyber security and is the organizer and host of the Telecom Threat Intelligence Summit. HardenStance openly supports the work of key industry associations, organizations and SDOs including NetSecOPEN, AMTSO, The GSM Association, MEF, OASIS, ETSI. The Cyber Threat Alliance. HardenStance is also a recognized Cyber Threat Alliance 'Champion'. www.hardenstance.com

HardenStance Disclaimer

special, indirect or similar damages (including lost profits), even if HardenStance Ltd was advised of the possibility of the same.

The user of this report agrees that there is zero liability of HardenStance Ltd and its employees arising out of any kind of legal claim (whether in contract, tort or otherwise) arising in relation to the contents of this report.