United States House of Representatives
Committee on Homeland Security
Subcommittee on Cybersecurity and Infrastructure

Statement for the Record
Gareth Maclachlan
Chief Product Officer
Trellix

"Security to Model: Securing Artificial Intelligence to Strengthen Cybersecurity"
June 12, 2025

Good morning, Chairman Garbarino and ranking Member Swalwell, and distinguished members of the committee. Thank you for the opportunity to testify today. I am Gareth Maclachlan, Chief Product Officer of Trellix. I have the honor of leading the team responsible for Trellix's product development, innovation and intelligence. I have been with Trellix since its formation, leading the divestiture and integration of McAfee Enterprise and FireEye to form one of the top five global cybersecurity vendors, and have held various roles in the company, and was previously general manager for the Network and Email security business. I have been in the cybersecurity industry for more than 25 years, with product and strategy roles at Mandiant / FireEye across cloud security, threat intelligence, incident response and managed detection services. I founded AdaptiveMobile, a cybersecurity vendor focused on telco and service provider security, and previously held roles in UK Intelligence agencies.

My testimony will address how our company is leveraging Artificial Intelligence (AI) to protect our customers from nation-state and cybercriminal cybersecurity threats; the evolving threats posed by AI; the promise we see for AI tools to help with cyber defense, especially with new iterations of AI; how we apply Secure by Design principles to the AI tools we use internally and in our product designs; how we address security baselines for AI and how how we monitor security after deployment; how we are working with the government on secure AI; and what role the Cybersecurity Infrastructure Security Agency (CISA) should play in the adoption of AI security.

Trellix commitment to cybersecurity
Trellix is the result of the merger of McAfee Enterprise and FireEye and is today a leading provider of enterprise-class cybersecurity solutions for the public and private sectors. CISOs worldwide trust our industry-leading products built on the broadest AI-powered security platform to help secure their organizations from advanced threats and strengthen operational resilience. Along with a large partner ecosystem, including leading providers of GenAI models, we focus on improving cybersecurity outcomes through AI, automation, analytics, and threat intelligence for more than 53,000 customers with responsibly architected security solutions delivered via cloud, on-premise, hybrid, and airgapped deployments.

Among our global customers, 50 percent of our business is with governments, critical infrastructure, and regulated industries. The US government is our single largest customer. We take pride in having served as a leading supplier to US federal agencies for decades. We are also active in voluntary federal cybersecurity public-private partnerships like the Enduring Security Framework and the Joint Cyber Defense Collaborative, given our commitment to protecting national interest.

Leveraging AI to protect customers from nation-state cybersecurity threats
While Trellix has used machine learning and other artificial intelligence capabilities to deliver leading-edge detection for more than a decade, Generative Artificial Intelligence (GenAI) presents a transformative yet complex landscape for enterprise security. While GenAI offers unprecedented opportunities for innovation and efficiency, it does so while challenging existing methodologies for product development and testing, and opens up new vectors for attack.

GenAI will fundamentally shift the economics of potentially every business process, removing the constraints of resource scarcity, prior experience, and developed skill. In every area, we as business leaders are implementing new strategies to take advantage of this freedom. But in parallel, we have to determine how to change control frameworks to adjust to AI innovations. Having one thousand additional GenAI software engineers sounds like a dream for any company. But constraining them to always deliver secure, reliable, scalable code consistently demands new procedures and inspection steps.

Outlining the changing risks: GenAI in the hands of threat actors
Cybercrime is a global business that seeks to exploit changing economics**.** Our threat intelligence organization monitors for indications of successful use of GenAI to change or expand techniques used by malicious actors from nation-states to financially motivated criminals to opportunistic attacks.

GenAI increases threat volume:

- **Deepfakes:** While already in use to create simulated material, we see the primary risk being to discredit or blackmail individuals or organizations, but doesn't by itself drive new security measures.
- **Phishing:** GenAI can significantly increase the scale and volume of phishing emails that can be generated, and make it easier to produce emails that would pass as coming from a native speaker. However, the purpose of any phishing attack is to get a user to do something they shouldn't–click a link, open a file, walk down the road to buy a gift card–our focus remains on the call to action, not determining whether or not a campaign is AI generated.

The greater impact will come from using GenAI to increase the sophistication of threats, and generate variability in the tools and attacks used:

- **Tool generation:** The greatest impact we expect in the short term is that the set-up and execution of an attack becomes significantly reduced, as threat actors use code-authoring and deployment tools to accelerate modifications to attacker capabilities, command and control platforms, and malware hosting. Rather than being constrained by hours in the day, a threat actor can use AI acting autonomously (agentic AI) to run multiple simultaneous attacks, all uniquely different against multiple victims in parallel, bringing organizations that may have had limited value into scope. If it costs 10 percent less to stage a ransomware attack now, a threat actor can pursue victims with a lower payout profitably–and even have the AgenticAI handle the ransomware negotiations.
- **Zero-day discovery:** Zero-days (never before seen exploitable vulnerabilities in a software package) have resurfaced as a leading reason for compromise. Typically, zero-days take significant effort from highly-expert researchers, and when found can be monetized among threat actors. Initial research indicates that GenAI models can be used to find new zero-days, raising the risk of rapid expansion in new vulnerabilities, but also enabling software companies to red-team their own products more exhaustively.

- **Cat-phishing:** Existing spam campaigns tend to be ineffective because they are a single mail or message. But with agentic AI, social engineering campaigns can be extended over time, and used to build a rapport and trust with every individual victim, at scale. While you might ignore the first email asking you to open an invoice pdf from an unknown supplier, it becomes a lot harder when that email comes after three months chatting about friends, family and how funny it was that you both happened to visit Arizona last summer.

Changing techniques are just one element: when a threat actor can employ transient agents running in the cloud, deployed, and monitored by other agents, the potential complexity for investigation and criminal proceedings by law enforcement will bring other challenges.

Novel threats from the use of GenAI by business
While threat actors will adopt GenAI to improve their economics, GenAI models being used legitimately by businesses add new attack vectors to be identified, monitored and secured. These are becoming well documented through existing frameworks such as NIST AI RMF, OWASP, and MITRE. Any organization leveraging available models for their own internal operations or for customer-facing services should evaluate the risk profile of all elements, namely:

- **Prompts:** Prompt hijacking and manipulation engenders unexpected outputs, thus limiting prompt variance by design while validating semantic intent and user-provided inputs such as links and files is critical. A prompt should be treated as potentially malicious code.
- **Responses:** The output from AI models, particularly those trained on company internal data, should be considered a potential vehicle for data exfiltration, and validated for PII and sensitive data leakage, exactly as a company would treat email. For organizations that need to consider restricted access policies, ensuring separation of content in training or filtering at response will be critical.
- **Models and training sets:** The integrity of company data used in training or to augment third-party models should be assured, with validation of new material and model performance, against baselines; and standard security measures for any enterprise application including encryption at rest and in transit, as well as robust access controls are necessary.

Bringing back balance: Using GenAI to accelerate security operations for enterprises
Starting more than two years ago, Trellix was the first company to use GenAI's deep, general security knowledge to understand security situations and decide on the best course of action to deliver effective cybersecurity protection outcomes. Unlike traditional software development where the same inputs will deliver the same outputs consistently regardless, the same is not true of GenAI–even when a model is isolated and a training set has not changed. A minor variation in a prompt such as an extra punctuation mark can cause a significant shift in response and, unlike in normal software where unexpected input would cause an error, detecting hallucination is challenging with GenAI models.

Our primary objective was to enable companies struggling with limited security operations expertise to gain the capabilities of nation-state defenders without the cost: an objective that couldn’t be met if human expertise was required to vet the output of every GenAI response.

We found several keys to success in leveraging GenAI for security operations:

1. **Guidance framework:** We instill a rigid guidance framework into our security architecture to prevent the AI from hallucinating or creating an output that is incorrect or misleading. This guidance framework was created from our own training set built up from hundreds of thousands of security investigations carried out by practitioners over the last ten years–i.e., using our own historical operations to constrain and validate output. Our guidance framework allows us to take advantage of the capabilities of the models without creating risk or sacrificing oversight.
2. **Unfiltered data:** We provide GenAI agents with unfiltered data, security events across network, email, endpoint, and cloud, which ensure the AI has information to make evidence-based decisions and not hallucinate. GenAI is effective at understanding new event formats, and decoding complex machine protocols that are not easily human readable, and we also use it now to auto-configure new feeds to add to the available data set.
3. **AI debates:** In product runtime, we require the AI to challenge its own decisions, defend its opinions, and prove its case with evidence. We do this by having "AI debates" in which different AI personas present their opinion and defend their position, then have an ultimate AI arbiter use our provided framework to choose the best position.
4. **Transparency:** At all times, we provide our users with full transparency into the decisions AI makes. This means showing the considered evidence and full analysis the AI provided, even when it decides no action is required. Humans ultimately must be able to understand and explain the basis for decision-making in AI.

As a result, we have successfully deployed our GenAI-enhanced security platform with multiple organizations globally, in many cases providing the capacity to ensure continual reassessment of every event as context changes, rather than having been limited by resourcing. Today,

- **Upskilling talent:** The analysis and decisions created by our AI are robust enough that customers use them to train beginner analysts into intermediate analysts, showing a glimpse into what the future holds for AI-human cooperation. This is especially critical as the skills shortage for cybersecurity practitioners grows.
- **Efficacious analysis:** We typically hear that AI outperforms humans at security event analysis, especially for the thoroughness of its investigations, as it will not get bored or distracted when trawling through thousands of events. AI excels at searching fields of haystacks for needles, straw by individual straw.
- **Hypothesis testing:** We use our extensive intel to provide threat models and attack paths to the AI which it interprets as hypotheses. It then uses agents to investigate and prove or disprove the hypothesis that the specific adversary is in the environment. It does this constantly in the background with ambient agents.

Attackers today are difficult to detect. Their activity rarely triggers an alert, but can be seen in a combination of low-level “informational” events. Without an AI platform, no human-only security operations team will ever have the bandwidth to isolate and investigate the transient indicators in time from the millions of similar but benign events.

Our customers have likened the deployment of our AI-powered Trellix Wise capabilities to increasing their security operations team 10-fold, while reducing variability and minimizing the impact of employee attrition and changing economics.

Applying Secure by Design principles to the use of GenAI in software development
As a cybersecurity company providing software to governments globally, Secure by Design methodology has been core to our processes. In addition to the steps outlined in the prior

section on using GenAI for effective and accurate responses, we have also adopted the following steps in our software development processes:

- **Prompt inspection:** In addition to constraining prompts, we also validate all files, URLs and other variables introduced by a user through our security analysis engines. We treat every question as a potential attack.
- **AI auditing:** We employ rigorous QA frameworks that use a combination of human-in-the-loop validation with AI-judging-AI to test and validate the behavior of different prompts (guidance frameworks/data) and models. We have published recent examples that show which models can be trusted with important decisions and when models hallucinate. Security after deployment depends on control and transparency. Our approach organically encourages the auditing of AI-based security decisions through its easily understood reasoning, as evidenced by our customers learning from and studying its analysis. We have both passive and active forms of human-provided guidance that customers use. Our AI will study the decisions humans make to use as guidance, and it will also listen to customer-specific instructions provided which gives it answers when it arrives at non-deterministic (coin-flip) situations.

Critically, while Trellix does train its analytics and machine learning models, we have elected not to train our GenAI models for two reasons:

- **Limit bias:** Fine-tuning models creates bias, causing them to place too much weight on previous situations and ignores their strength in general security knowledge.
- **Privacy concerns:** It introduces privacy concerns in which data or decisions made in one customer environment could impact an unrelated customer.

We are also evaluating the use of GenAI tools to accelerate product engineering internally, expanding our software output capacity. In accordance with Secure by Design principles, we have implemented key controls:

- **Third-party models:** We leverage only third-party models that are isolated from the provider and hosting service, and do not leverage our content in model training.
- **Code evaluation:** We treat all code generated by a GenAI coding engine as suspect, and leverage agenticAI models to peer review, evaluate for performance and security.
- **Test coverage:** We use GenAI to generate extensive test coverage for all our products - the test plans are evaluated and validated by our teams, but ultimately use of AI in QA leads to more secure, more robust and performant code.
- **Red-teaming:** We are evaluating the potential for GenAI agents to provide exhaustive and continual red-teaming against our products.

As the industry moves forward, our perspective is that the validation and attestation of models will become increasingly important, and that companies seeking to reduce risk may increasingly move away from hosted LLMs. For example, working with governments, defense departments and regulated industries, we are demonstrating the benefits of moving to closed small language and quantized models:

- Reducing the variability in output that comes from LLMs
- Providing a known baseline that is invariant for testing subsequent model updates
- Ensuring sensitive prompts and response are airgapped
- Improved economics

How we partner with CISA to secure AI
We are working directly with CISA, as well as with other agencies and branches, to enable defenders to keep pace with attackers and maintain a secure future by establishing organizational goals and policies into the strategic and tactical guidance frameworks AI needs for safe, automated defenses. In particular, we are sharing our expertise in using GenAI to select from 1,000+ investigative actions, built from historical analyst use of our platform, to constrain decisions and prevent hallucination; our advanced frameworks for measuring the accuracy of GenAI decisions and activity against baselines; and our experience using GenAI to ingest and adapt to new data sources to accelerate detection engineering. This approach delivers on the efficiency promise of GenAI, while preserving the transparency and explainability of decision making, important when moving towards the goal of automation of response actions.

The Secure by Design attestation framework championed by CISA is a positive example of the role CISA has for the industry: the approach encouraged companies like Trellix to review existing processes, to verify internal compliance and controls were operational, and to reduce risk for both supplier and customer; benefitting not just US Government departments, but all enterprises globally. We suggest CISA take a similar role and provide a framework model that providers should adopt and attest, permitting any enterprise seeking to adopt GenAI can do so with confidence in the efficacy, safety, and resilience of available commercial and open source models. By continuing to leverage a voluntary public-private partnership model, CISA can use its convening power to help the private sector gain the benefits of AI security while mitigating its risks.

Mitigating AI risk for future advancement
The constant advancement of adversaries and their use of AI makes defenders' use of AI compulsory. Trellix, along with our partners and peers in the cybersecurity industry, are leveraging advances in AI to enable our customers to protect themselves from nation-state and global cybercriminal enterprises; across all aspects of detection, security operations and product development. There is significant risk when allowing AI to make decisions in automation, but the right guidance frameworks and expertise can mitigate these risks.

Organizations adopting GenAI models need to be able to do so with confidence in the development lifecycle of model providers, continuing transparency in model outputs, bias, and risk from providers, and with industry guidance on the security controls that can reduce the introduction of new vectors of attack.