CYCRAFT WHITE PAPER

AI Adopting Governance: From AI-Enabled Security, AI-Native Safety to Autonomous Exposure Management

2026 AI Adopting Whitepaper

2026.03

To implement Autonomous Exposure Management, CyCraft advocates a foundation built on an AI Security Platform (AISP), integrating Unified Exposure Management (UEM). This approach leverages AI automation to handle current cybersecurity issues while further utilizing AI to address AI-native and potential LLM threats. This white paper proposes an actionable framework for introducing AI into cybersecurity, demonstrating how CyCraft's XCockpit and XecGuard empower enterprises across different application scenarios.

2026 Technical Trend: AI Adopting Security Governance

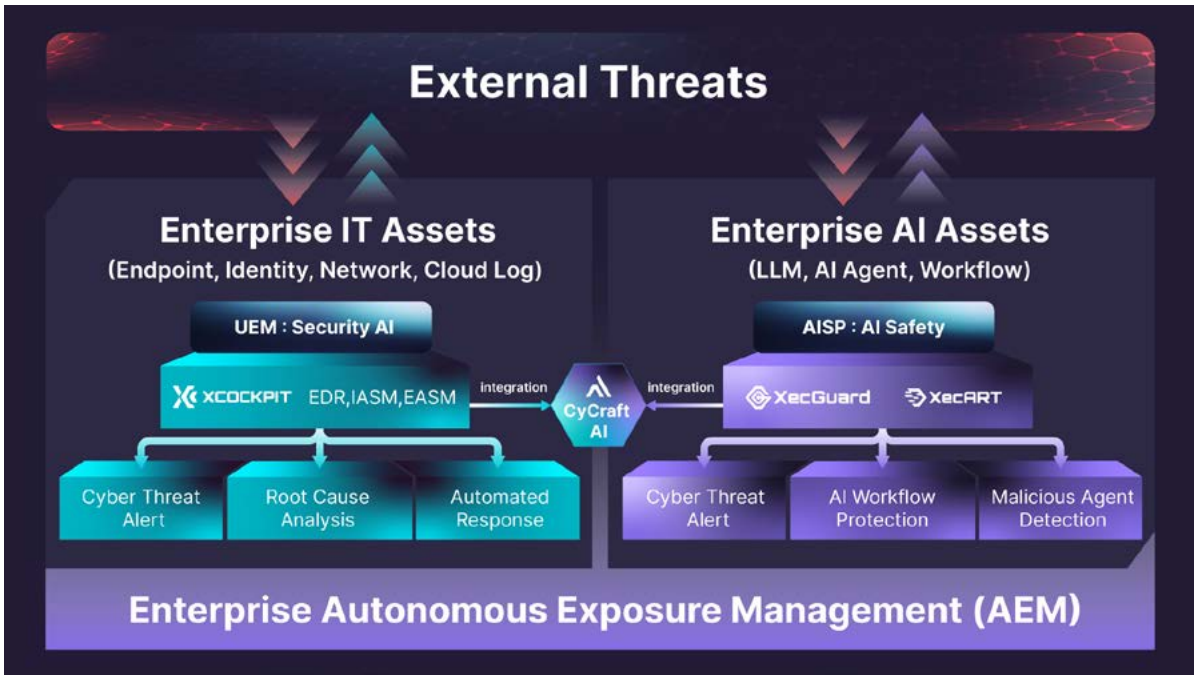
Since the launch of ChatGPT in late 2022, the volume and scale of attacks have continuously broken historical records. It is predicted that global cybercrime costs will grow at a rate of 15% annually. If the economic scale of cybercrime (approximately $10.5 trillion by 2025) were considered a country, it would be the world's third-largest economy, trailing only the United States and China. Even for social engineering attacks, the cumulative increase in phishing emails has reached 4,151% (statistics up to mid-2024). Because AI-generated content is more customized, traditional defense false negative rates have risen significantly. Attackers also benefit from AI automation technologies, scanning enterprise network vulnerabilities 24/7 and combining them with supply chain attacks. This has led to a sharp increase in the average number of weekly cyberattacks suffered by global organizations, growing from 1,168 in 2022 to 2,003 in 2025—an increase of 71% (statistics up to November 2025).

AI-driven emerging cybersecurity risks challenge traditional defense strategies while simultaneously catalyzing the application of various AI technologies and security mechanisms. Security service providers who have adapted to this paradigm shift are beginning to use AI automation to handle existing issues and are now facing the potential security problems of AI-native and LLM models. Although these two aspects differ in the depth and breadth in terms of AI automation, they belong to the same defense spectrum, giving rise to the AI Security Platform (hereinafter referred to as AISP).

AISP integrates disparate security tools to comprehensively and continuously protect models, data, LLM applications, and AI posture management with AI at the core. It assists security teams in decision-making and risk prioritization from an overall perspective. Since traditional security tools mostly focus on endpoints, identity authentication, and network risks, they not only fail to grasp the AI development process but also lack security testing for prompt vulnerabilities and protection mechanisms against malicious AI Agents. AISP comprises two main pillars: AI Usage Control and AI Application Cybersecurity. The former manages regulations for enterprise employees and systems using third-party AI services, while the latter extends protection to customized AI applications such as enterprise-built LLMs, fine-tuned models, and AI Agents.

We believe that a comprehensive and autonomous security platform must incorporate AI application governance; these two should not be managed separately. Through this unified management platform, all critical components and actions of the AI system are continuously monitored. Scattered alerts are transformed into actionable threat intelligence. When simulating attack paths or automatically correlating case details, the platform rapidly generates management measures that are compliant and conducive to subsequent decision-making. The cybersecurity management system in the AI era should be based on AISP and incorporate the concept of Unified Exposure Management (UEM) to fully utilize AI automation technology and implement Preemptive Exposure Management (PEM) security mechanisms.

> ***A comprehensive and autonomous security platform must incorporate AI application governance; these two should not be managed separately.***



Unified Exposure Management: AI-Enabled Automation Targets Long-Term Issues

In the AI era, traditional cybersecurity issues persist and have become more severe. The main problems we observe include:

- Lag in Threat Detection and Insufficient Response: Traditional signature-based defense methods can effectively defend against known threats but cannot cope with unknown vulnerabilities such as zero-day attacks or new malware variants.
- Data Explosion and Indigestible Intelligence: While SIEM platforms or SOC teams collect massive amounts of data, they lack tools to deeply analyze intelligence from different sources and cannot effectively correlate it with internal alert data, rendering the information useless.
- Increased Scale and Complexity of Attack Surfaces: The rise of cloud computing, mobile devices, and remote working pattern has shattered traditional network perimeters. Firewalls and VPNs are unable to protect scattered assets. Furthermore, internal malicious acts or human errors within enterprises are difficult to capture through traditional network traffic analysis, leading to a lack of visibility into exposed attack surfaces.
- Shortage of Cybersecurity Talent: Enterprises struggle to recruit or retain talent capable of threat anticipation and incident response, creating gaps in the defense system. To make matters worse, a high volume of false positives and alert fatigue stretch manpower even thinner, making it easier to miss genuine major threats.

Automation has accelerated the fragmentation and volume of information, turning the scarcity of manpower and resources into a larger breach. However, the crisis brought by AI technology is also a turning point. As AI brings scale to cybersecurity capabilities and human intelligence, many service providers have begun to integrate AI to address existing cybersecurity problems.

"Rome wasn't built in a day." The key to breaking the deadlock with AI or machine learning lies in shifting from passive defense to proactive cybersecurity. UEM provides a unified, comprehensive view of exposure, prioritizing various vulnerabilities and assets based on risks. This allows security teams to stop firefighting and dismantle threats in a structured, sequential manner.

Beyond vulnerability management, UEM emphasizes attack path analysis, external attack surface management, security configuration management, and identity and access management. By effectively integrating vulnerabilities, configuration, identities, and assets onto a single platform, UEM transforms security mechanisms from "patching holes" into "plannable fortifications."

We believe a security system capable of realizing the UEM concept should possess at least the following functions:

- Behavioral Anomaly Analysis: Distinct from traditional signature or malicious file detection modes, this establishes a database of user and entity behavior patterns to automatically identify behaviors that deviate from the norm, detecting attacks, insider threats, or fileless malware in real-time.
- Automated Root-Cause Analysis: Shifting from passive response to active analysis. It automatically simulates attack paths, analyzes malware samples, and extracts the potential origin, propagation method, and impact scope of attacks from billions of logs, providing concise incident briefing to bridge the cybersecurity talent gap.
- Context-Oriented Alert Analysis: Consolidating large and complex alerts into single cases, visually presenting event timelines and correlations. Adding potential background information to effectively reduce noise can lower the cost of manual judgment and the false positive rate.
- Predictive Risk Scoring: Integrating external threat intelligence (e.g., surface web, dark web, leaked enterprise credentials), internal asset importance (e.g., asset risk level, category, permissions), and vulnerability severity. Through structural and component identification analysis, it correlates the attack surface of accounts, endpoints, and services to predict which weaknesses are most likely to be exploited by attackers, automatically suggesting the optimal remediation.
- Autonomous Response and Remediation: An AI-driven platform can automatically execute preset response playbooks based on the detected threat type (e.g., isolating infected endpoints, terminating malicious processes, blocking IPs), create tickets for high-severity events, track mitigation measures, and record various risk handling statuses and performance.

Application Scenario: CyCraft EASM Consolidates Intelligence, Digital Assets, and Risk Prioritization

Incorporating AI automation technology into existing security tools empowers organizations of leaner security manpower and resources. Taking the government sector as an example, Class A and Class B government agencies regulated by the *Cyber Security Management Act* bear a heavy cybersecurity burden due to their involvement in critical national development, core technology research, critical infrastructure, and shared system operations. They are required to regularly audit the security performance of subordinate units. However, the nature of these units varies, and the number of subordinate organizations—including public sectors, state-owned enterprises, and legal entities—is vast and diverse. Most units are accustomed to using static questionnaires, sending and collecting data one by one. This is a labor-intensive and time-consuming process, since significant manpower and time are spent organizing, selecting, archiving, and managing obtained data.

To address this scenario, adopting the UEM framework allows the direct acquisition of endpoint and site information from subordinate units, simplifying management and audit processes, reducing human error, and increasing information credibility. The CyCraft EASM (External Attack Surface Management) module can provide concrete and actionable risk assessment indicators, audit recommendations, and control measures tailored to the different security management systems and regulatory requirements of each site. This enables supervisors or colleagues without any cybersecurity background to understand the implications, satisfying cross-departmental collaboration needs and reducing communication costs.

CyCraft EASM integrates intelligence and asset trajectories, automating attack path prediction and risk assessment.

Preemptive Exposure Management: Defend AI-Native Safety via AI Technology

While AI-driven defense tools can resolve legacy issues, the risks inherent in AI and LLMs also introduce emerging threats, and Agentic AI makes management even more difficult. The uncontrollable nature of LLMs means enterprises cannot confirm details before adopting AI tools. Other use cases are not reliable references because differences in application scenarios or industries, and issues that arise after implementation cannot be directly fixed, creating a management crisis. Attack methods such as Prompt Injection, Prompt Extraction, and Jailbreak can trick LLMs into following attacker commands, leaking training data, or generating harmful content, turning AI into an accomplice for malicious behavior.

In the era of Agentic AI, attackers can use Prompt Injection to make AI Agents autonomously execute malicious code, send phishing emails to customers, or delete database records, resulting in wider impact and harder control. Since AI Agents are granted permission to access external resources (e.g., Email APIs, databases, code execution environments), malicious prompts can trick AI Agents into abusing external tools, transforming language commands into unauthorized operations. Once a malicious prompt is successfully injected, the AI Agent may decompose this malicious goal into multiple sub-tasks and execute them over a long period, making it difficult for humans to detect and block in real-time.

Recently, researchers have noted the rise of the Promptware family, which uses the victim's AI environment as infrastructure to generate malicious behavior in real-time. AI is not only used for development but has become an operational component of attack tools. If enterprises do not thoroughly inventory and manage internal AI applications, exposed access points will become breaches for attackers to exploit "living off the land." Additionally, since AI generates malicious commands or scripts in real-time during execution, static signatures are drastically reduced, rendering traditional signature-based detection methods (like YARA) ineffective and allowing attackers to further evade monitoring and analysis tools.

Currently active Promptware families include FRUITSHELL, PROMPTFLUX, and PROMPTLOCK.

- FRUITSHELL embeds adversarial prompts within code, not just using AI to assist attacks but using prompts specifically to deceive AI security tools.
- PROMPTFLUX treats LLMs as dynamic code for obfuscation, self-morphing during the execution phase to evade detection.
- PROMPTLOCK uses the victim's local LLM as a ransomware script engine, generating and modifying attack scripts directly on the endpoint in real-time.

Whether dealing with risks latent in AI/LLMs or direct prompt abuse of AI systems, new threats are emerging endlessly. We must move beyond proactive defense towards Preemptive Exposure Management (PEM) mechanisms. PEM is one of the top ten strategic technology trends for 2026, emphasizing the use of AI or machine learning to deeply analyze massive amounts of data to rapidly provide credible attack predictions and resolution measures. As Gartner VP Analyst Tori Paulman stated: "These trends represent more than technology shifts; they are catalysts for business transformation… Because the next wave of innovation isn't years away, organizations that act now will not only weather volatility but shape their industries for decades to come."

To adapt this paradigm shift, the core is to use AI to monitor and defend AI—denying, deceiving, and disrupting attackers before they complete an attack. This involves deploying a plan to dismantle the kill chain in advance, rather than just optimizing alerts or post-incident response. Compared to proactive security, which emphasizes taking action before alerts are triggered (such as threat hunting, purple teaming, and adversarial simulation), PEM focuses more on remodeling the environment to cause attacks to fail early or become economically unviable. Reducing the potential attack surface, increasing the cost required for a successful attack, and controlling the scale of events so they do not evolve into incidents are the operational essentials of this framework.

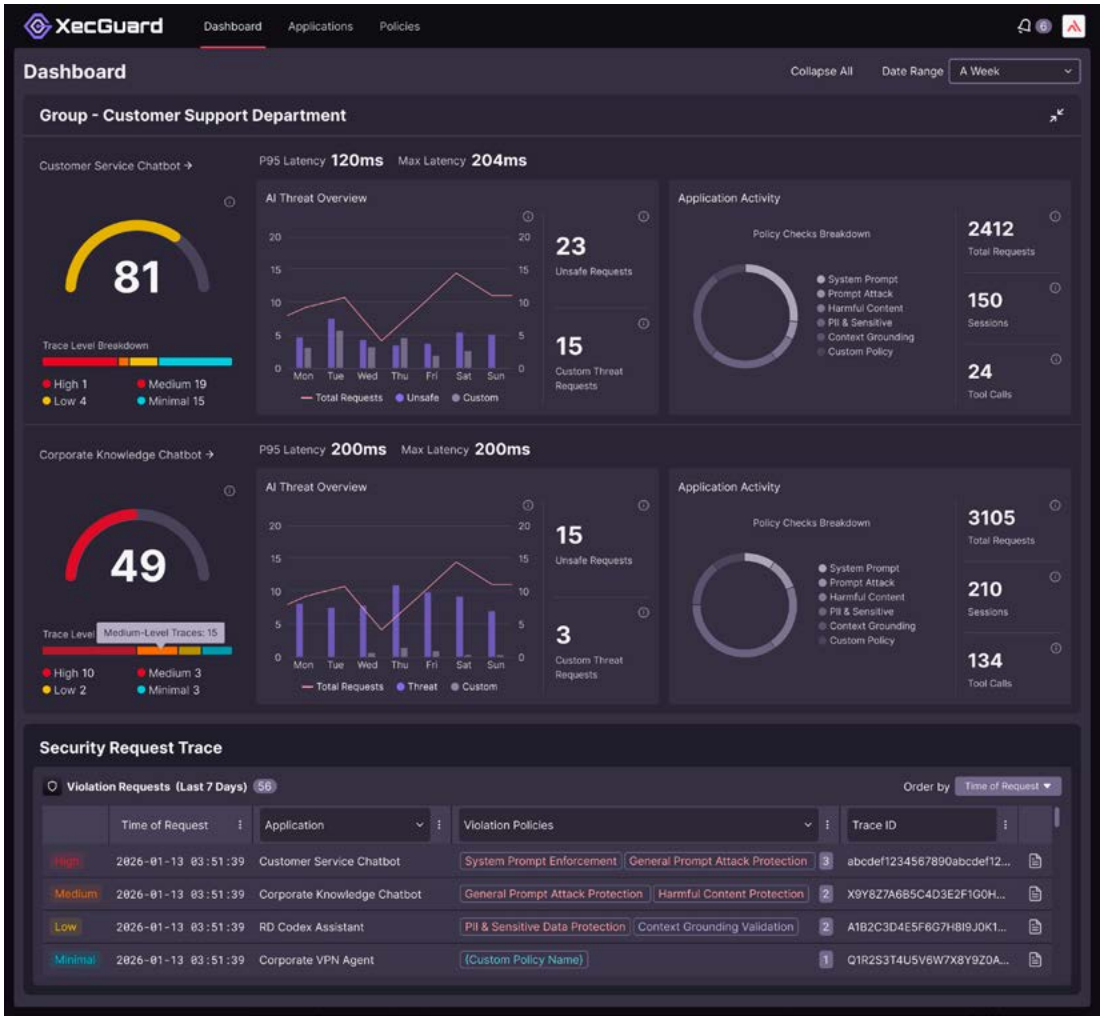
In 2025, CyCraft was twice selected as the Gartner sample vendor. We are a pioneer in applying PEM and UEM, being the only benchmark enterprise from Taiwan. When designing solutions, we believe that focusing on AI-native safety is the primary way to truly build a PEM platform. Through the following features, we assist users in making AI and LLMs more robust:

- Zero-Code Safety Hardening: Directly introduces security protection via API interfaces compatible with common LLMs. It requires no code changes, seamlessly integrates with existing AI applications, and can block Prompt Injection attacks, check for sensitive PII, manage prompts, and audit LLMs.
- Flexible Deployment of AI Guardrails: This provides both passive detection and active blocking modes for different scenarios such as on-premise or third-party applications, with flexible deployment to minimize interference with AI performance. For AI Agents, it provides retrieval guardrails to ensure the credibility and security of scenarios like RAG and function calling.
- Scalable AI Safety Governance: This features a cloud-native auto-scaling framework with multi-tenant management, auditing, and alerting capabilities. Enterprises can formulate independent policies for different sites and integrate with existing SIEM platforms or SOC teams via CEF Syslog and Alert APIs to create a complete and consistent AI security governance system.

Application Scenario: CyCraft XecGuard Integrates Industry Regulations to Solidify AI-Native Safety

Today, AI applications are highly integrated into enterprise environments, making AI-native security crucial for enterprise security management. Take the financial industry as an example: since this industry serves the public and has complex operations, there is a strong demand for AI applications. However, personal privacy and transaction information are highly sensitive. The Financial Supervisory Commission and related financial departments have established numerous regulations, making the use of on-premise or open-source models difficult. Without perfect AI usage control or application security standards, problems such as Shadow AI, information leakage, supply chain risks, and intellectual property disputes are highly likely to occur.

CyCraft XecGuard, the next-generation AI firewall, fuses red and blue team experience in critical fields such as government, finance, healthcare, and high-tech manufacturing. It focuses not only on the defense upgrade of the model itself but also integrates financial regulations and management rules to avoid abuse and prevent derivative legal or ethical issues. It assists enterprises in ensuring that information security, regulatory compliance, and system resilience are upgraded synchronously while quickly deploying AI. XecGuard is a solid line of defense that strengthens AI with AI.



CyCraft XecGuard effectively protects LLMs from multiple Prompt attacks.

AI Adopting Security Overview: From Development, Deployment to Risk Management

Leveraging AI to integrate and process multi-modal exposure information and incorporating the concept of PEM into Autonomous Exposure Management is what we consider a forward-looking AI security solution. To ensure the safety of AI applications, concepts such as AI usage control and prompt management should be included in all stages of its lifecycle. Prompts should not be viewed merely as text narratives, but as the executable specifications and strategy layer of the AI system. In AI applications, their importance and influence are equivalent to the source code of traditional systems.

We recommend that during the AI application system development stage, AI applications and system prompts should undergo prompt quality assessment and automated review, just like source code, to comply with company security policies. AI development should follow the AI Lifecycle Development Cycle; prompts must be versioned, audited, and traceable. Any changes should automatically trigger the CI/CD pipeline, supplementing the deficiencies of the traditional SDLC.

> ***“Prompts should not be viewed merely as text narratives, but as the executable specifications and strategy layer of the AI system. Their importance and influence are equivalent to traditional source code.”***

In the process of AI security operations, given that Prompt Injection is listed by NIST as the highest risk for AI systems, we believe prompt attackers should be viewed as traditional hackers. Malicious behavior abusing prompts is ought to generate alerts and assessment reports for development and audit teams. Furthermore, enterprises need to establish guardrail policies to record all prompt communications between users and AI, and between AI Agents, to execute real-time and continuous monitoring. Security teams should also regularly conduct adversarial AI red teaming to evaluate AI chatbots.

At the AI risk management level, it is essential to simultaneously inventory approved AI applications and those actually used to avoid Shadow AI risks. An independent LLM should be used to audit all AI activities within the company to prevent model self-auditing bypasses. Prompt auditing must include internal communications between LLMs and between AI Agents, incorporating application scenarios such as RAG and function calling. AI guardrails capable of automatically analyzing prompt behavior are key to risk management and data protection.

In the future, across industries such as healthcare, high-tech manufacturing, ICT, and transportation, whether for internal cross-unit massive data exchange or for providing more personalized and complex services to the public, operations will rely on AI and LLMs as their core. Therefore, we propose an AI-centric integrated autonomous risk management platform, along with key operational points and solutions for different application scenarios. Empowering AI to defend and strengthen AI can ultimately elevate cybersecurity defense capabilities.

About CyCraft

CyCraft Technology (7823.TW) — Taiwan's first AI-native cybersecurity company to list on the TWSE Innovation Board — is dedicated to automating cybersecurity with AI, shifting defense from reactive firefighting to proactive, scalable dominance. Delivering protection that is faster, easier, safer — and smarter on cost, CyCraft empowers defenders to scale. With a proven track record serving top-tier government agencies, leading financial institutions, and semiconductor giants, CyCraft is building Asia's most advanced AI-driven joint defense ecosystem — dramatically shortening dwell time, reducing breach impact, and strengthening enterprise digital resilience worldwide.

www.cycraft.com


CYCRAFT