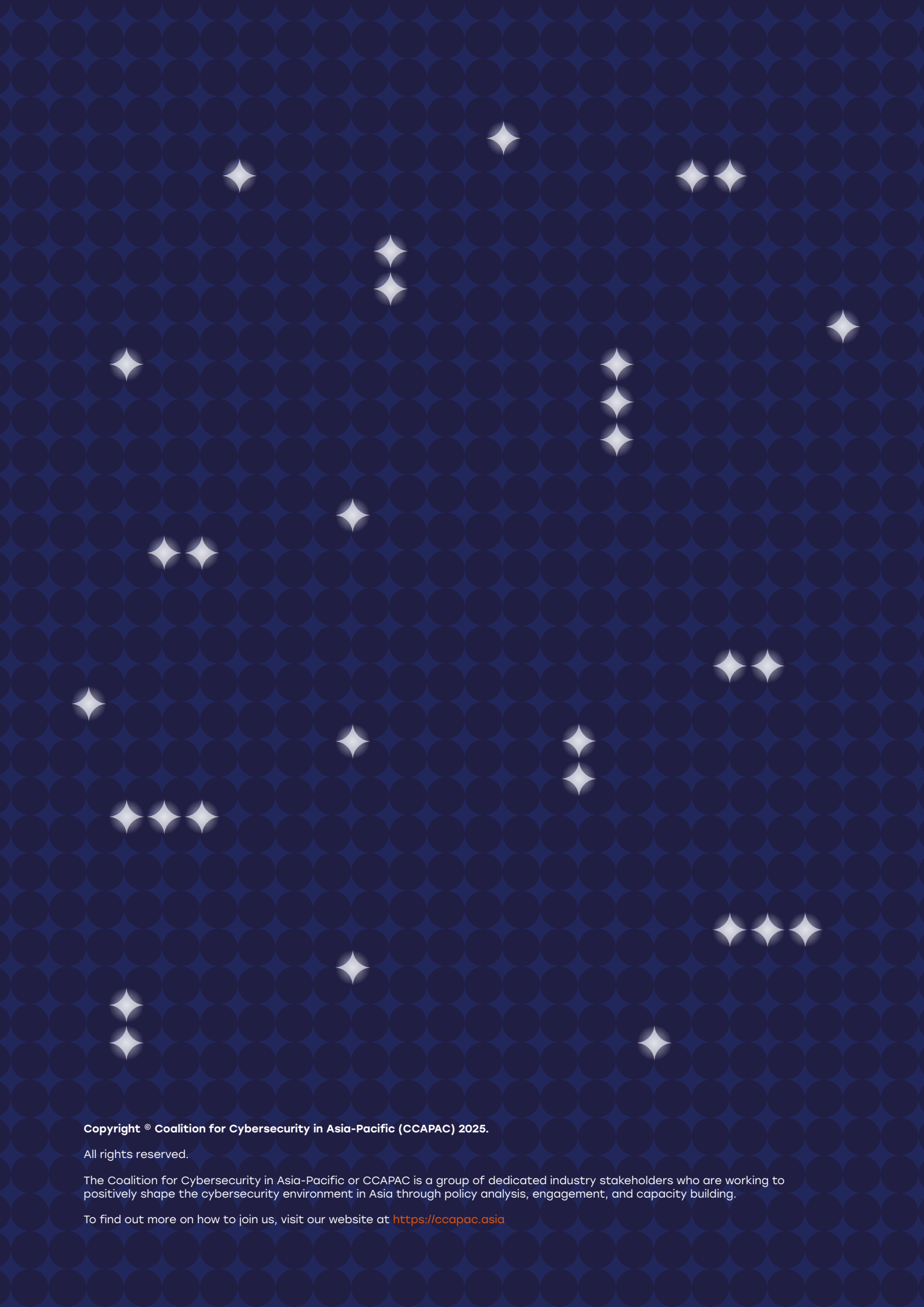
CCAPAC Annual Report 2025

AI Security in 2025 and Beyond: Emerging Threats and Solutions

COALITION FOR CYBERSECURITY IN ASIA-PACIFIC

ccapac.asia



The Coalition for Cybersecurity in Asia-Pacific or CCAPAC is a group of dedicated industry stakeholders who are working to positively shape the cybersecurity environment in Asia through policy analysis, engagement, and capacity building.

To find out more on how to join us, visit our website at https://ccapac.asia

Table of Contents

Executive Summary

The AI Security Challenge

Artificial intelligence has evolved from experimental technology into a mission-critical business function, with 78% of companies worldwide actively deploying AI across operations.[1] However, this rapid adoption has introduced new security challenges that demand urgent attention from security leaders and policymakers alike.

Two Critical Emerging Threats

The 2025 Coalition for Cybersecurity in Asia-Pacific (CCAPAC) report identifies two critical threat categories reshaping the cybersecurity landscape:

- **Agentic AI Security Risks:** Unlike conventional AI applications, agentic AI systems operate with autonomous problem-solving capabilities. These systems introduce distinctive threat vectors beyond traditional software vulnerabilities. A notable incident involved attackers weaponizing Claude Code,[2] an agentic AI coding assistant, to execute sophisticated data extortion operations across 17 organizations spanning multiple sectors.

- **AI-Powered Social Engineering:** Cybercriminals are industrializing phishing attacks through AI automation. Globally, organizations experienced a 17% surge[3] in phishing emails, while attackers have exploited AI-driven platforms like Lovable to generate over 5,000 malicious phishing URLs.[4] AI systems now outperform[5] elite human teams in phishing effectiveness, making attacks more persuasive, proficient, and prolific at an unprecedented scale.

Market Response and Solutions

The urgency of these threats is reflected in market dynamics: the AI security market is projected to surge[6] from USD 20.19 billion (2023) to USD 141.64 billion by 2032.

Emerging defense technologies include containment systems, multi-agent security frameworks, advanced human risk management, real-time synthetic media detection, and adaptive email security defenses.

Governance Gap

While governments across Asia-Pacific (APAC) and global jurisdictions have developed AI national strategies, security considerations remain nascent and/or inconsistent. Current regulations lag behind technological advancement, creating gaps deserving attention, particularly concerning agentic AI systems and AI-enabled phishing where governance frameworks are largely in their infancy.

Strategic Recommendations

To bridge the widening gap between AI innovation and security, this report recommends:

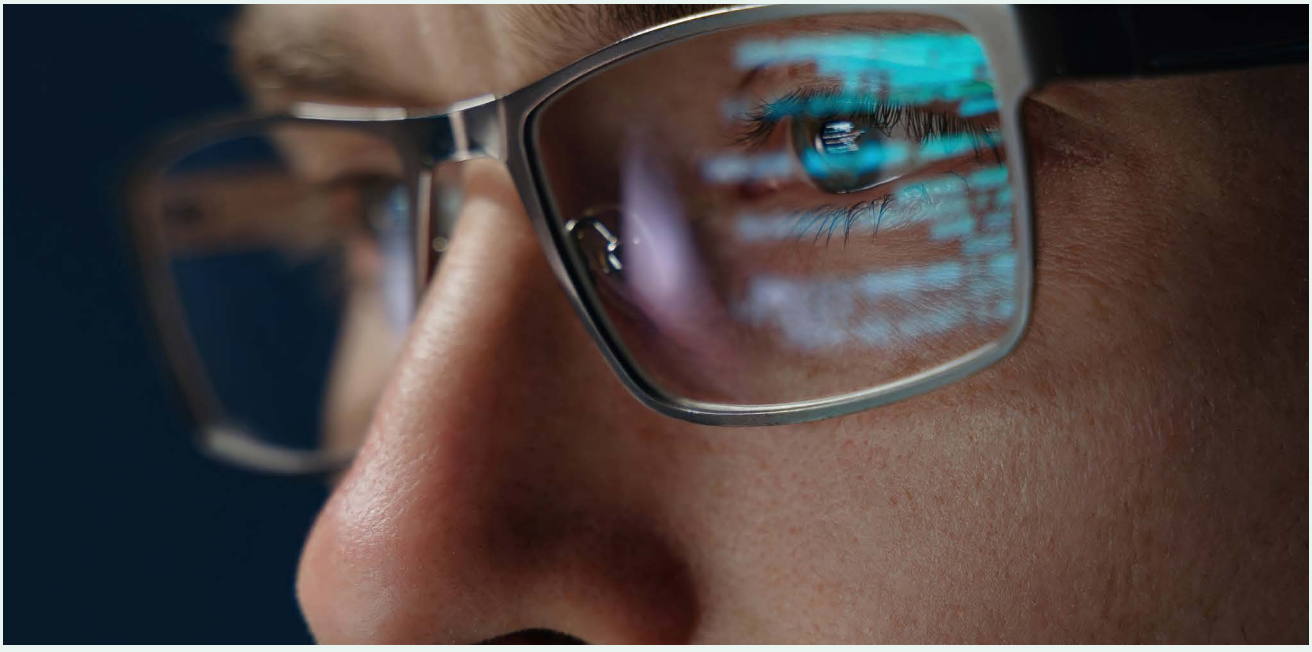
1	Establishing public-private partnerships on AI security threat intelligence
2	Strengthening dynamic, expert-driven AI security frameworks
3	Implementing regulatory sandboxes for AI security innovation
4	Promoting mutual recognition agreements for AI security certifications
5	Prioritizing security awareness and investing in AI cybersecurity skills development

The Path Forward

The transition to agentic AI demands careful recalibration of technical controls and governance structures. Similarly, the industrialization of AI-enabled social engineering has also created new risk vectors.

Success requires coordinated action across governments, industry, and research institutions to establish proportional, evidence-based safeguards that preserve innovation benefits while containing systemic risk.

The objective is not to constrain progress but to make accountable autonomy achievable through sustained investment in technical capabilities and evidence-based frameworks.



Chapter 1
Introduction

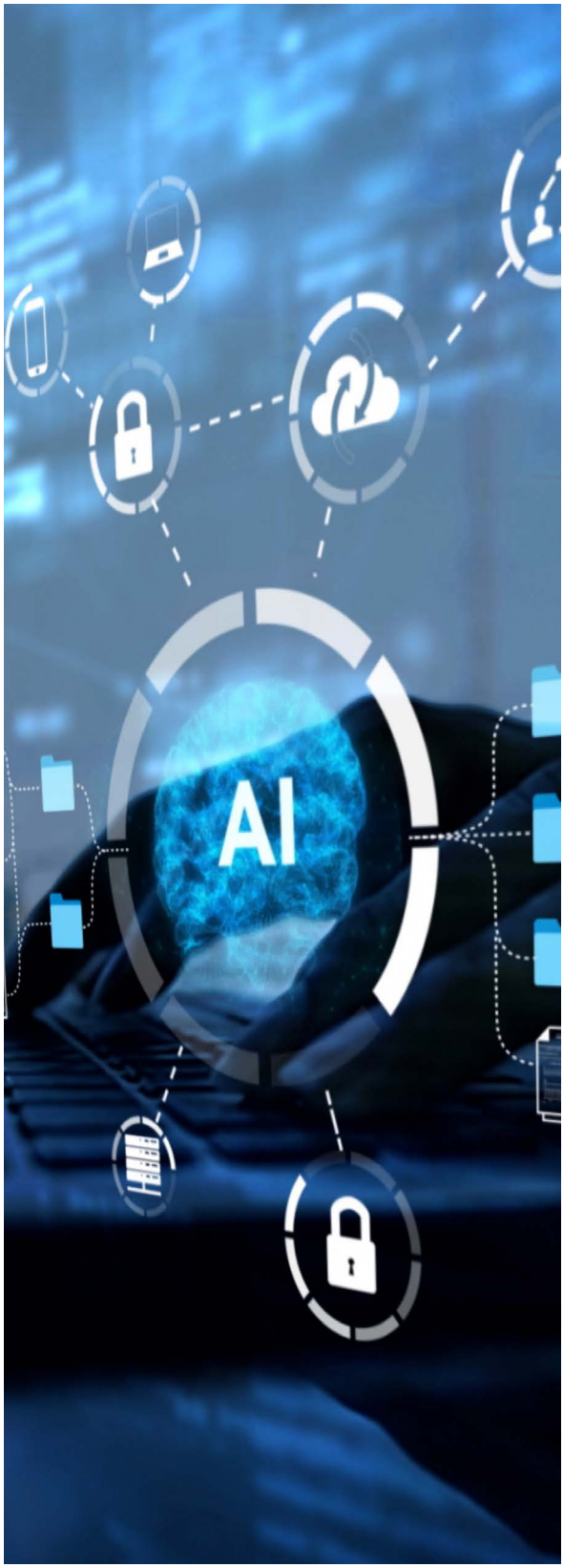
Artificial Intelligence (AI) has shifted rapidly from experimental technology to a central driver of global business operations. Today, 78%[7] of organizations report using AI in at least one business function, up from 72% earlier in 2024. Only 13%[8] of companies report having no plans for adoption, meaning nearly four in five organizations are engaging with AI in some form. On a global scale, AI is now deeply embedded in people's daily lives: 66%[9] of individuals globally report intentionally using AI with some regularity, while 56%[10] of citizens now believe that AI will positively transform their lives in the next 10 years.

This level of adoption reflects AI's broad advantages[11]: organizations use AI to accelerate decision-making, automate processes, personalize customer engagement, improve fraud detection, innovate in product development, and optimize logistics.

While the advantages are significant, AI deployment has also introduced new categories of security risk. Current concerns include model inversion attacks, data leakage, adversarial inputs, and vulnerabilities in AI application pipelines. These issues shaped the 2024 CCAPAC Report,[12] which provided a holistic overview of AI risks across data, models, infrastructure, and applications.

However, as AI technologies continue to evolve, so too does the threat landscape. This 2025 CCAPAC report focuses on two emerging areas that have come into sharper relief this year: **agentic AI security risks and AI-powered social engineering.** These developments, highlighted in the Cisco State of AI Security Report 2025[13] and the G7 Cyber Expert Group,[14] represent a new class of threats shaped by the increasing autonomy and contextual awareness of AI systems.

While these security issues might have seemed theoretical some months ago, reality has caught up. In August 2025, researchers at DEF CON, the world's largest and most famous hacker convention, demonstrated how customer-facing AI agents can be manipulated[15] into performing unauthorized actions, such as exposing internal tools and customer relationship management records.



At the same time, cybercriminals are using AI to scale social engineering campaigns. In October 2025, OpenAI revealed how threat actors exploited ChatGPT[16] to streamline phishing campaigns and malware distribution. The attackers followed a detailed playbook, crafting polite, professional emails that impersonated academics, industry figures, or conference organizers, and targeted specific demographics.

These developments illustrate a pivotal moment of convergence for AI and cybersecurity. While traditional risks remain, agentic AI vulnerabilities and AI-powered social engineering create an acceleration vector that is proving both tangible and urgent.

By focusing on emerging AI threats, the 2025 CCAPAC report will equip decision-makers with a clearer understanding of the AI security frontier and how government and industry can come together to respond through proactive collaboration, standards, and controls. Together with our colleagues in the public, private, and people sectors, CCAPAC continues to positively shape the Asia-Pacific (APAC) cybersecurity ecosystem through policy analysis, engagement, and capacity building.

This report is structured as follows:

Chapter 2

Reviews authoritative threat intelligence reports, highlighting the convergence between industry and government in identifying AI agents and AI-enabled social engineering as key emerging concerns;

Chapter 3

Shows how the security industry is reacting to these new threats, offering initial organizational-level solutions for early AI adopters;

Chapter 4

Gives an overview of AI security policy developments in 2025, focusing on selected APAC countries and western jurisdictions;

Chapter 5

Provides policy recommendations to governments across the APAC region and beyond on how to collaborate with responsible stakeholders to mitigate current and future AI security threats while reaping the benefits of this transformative technology.

Chapter 2
AI Security in 2025: Emerging Threats

The cybersecurity landscape is experiencing a fundamental transformation as AI becomes deeply embedded in organizational operations. Recent industry analyses on AI security highlight critical challenges[17] that demand urgent attention, including serious threats to AI models, systems, applications, and infrastructure; AI-specific attack vectors; and the use of generative AI to automate and professionalize threat actor cyber operations.

In 2025, leading industry and government threat intelligence reports–Check Point Software,[18] Cisco,[19] CrowdStrike,[20] Palo Alto Networks;[21] ENISA,[22] U.K. NCSC,[23] Singapore Cyber Landscape[24]–have consistently identified two key emerging threats that security professionals must monitor closely: the deployment of agentic AI systems and increasingly sophisticated AI-powered social engineering.

These two developments represent a critical inflection point where AI capabilities have reached sufficient sophistication to fundamentally reshape both offensive and defensive cybersecurity operations. This chapter delves into these two emerging AI security threats.

2.1 Agentic AI Security

The emergence of agentic AI represents a fundamental technological shift in how organizations approach automation and decision-making. Unlike traditional AI systems that respond to specific inputs with predetermined outputs, agentic AI combines large language models with reasoning capabilities and autonomous action-taking abilities.

Nonetheless, current adoption patterns reveal a significant security gap in early adopters: 82%[25] of surveyed organizations are already deploying AI agents, yet only 44% have established security policies specifically designed to counter threats stemming from their usage, leaving a substantial portion of these organizations operating without adequate protections.

Key Features

Agentic AI systems[26] distinguish themselves from conventional AI applications through their capacity for independent problem-solving within defined environments. This represents a significant shift[27] from systems that react to predefined inputs to those that can autonomously set goals, plan, and execute complex tasks.

Unlike traditional AI models that operate in isolation, agentic AI systems distribute tasks across autonomous agents that communicate and collaborate, enabling faster execution, continuous innovation, and autonomous self-healing networks. These systems operate through multiple integrated modules that synthesize information from various sensors and databases to establish contextual awareness for autonomous decision-making processes.

The defining characteristics of agentic AI encompass these core capabilities:

- autonomous operation independent of human oversight
- cognitive learning from previous decisions and environmental observations
- communicative sharing of operational states with surrounding systems
- modal adaptability (i.e., the ability to adjust and perform with different data inputs) based on environmental conditions
- proactive action initiation
- reactive responsiveness to environmental changes
- robust compensation for internal and external disturbances
- social communication capabilities with other AI agents

An example is Salesforce's Agentforce,[28] which deploys autonomous AI agents within Salesforce environments for sales, customer service, and operational functions. These agents demonstrate the practical application of agentic AI by proactively initiating workflows, analyzing real-time customer relationship management data, coordinating across multiple tools including Slack, drafting business proposals, qualifying sales leads, resolving customer cases, and executing decisions based on predefined business rules. None of these actions performed by AI agents require manual human intervention.

Threats

Agentic AI presents significant opportunities to enhance software and organizational security, for example by automating software vulnerability detection[29] and security operations.[30] AI agents enable accelerated threat detection by processing security alerts at speeds exceeding human analytical capabilities, while providing automated defense mechanisms through intelligent triage and scalable response systems. Multiple specialized AI agents can work collaboratively[31] to handle different aspects of security operations, with applications in network security, identity management, and incident response.

Scientific research found that autonomous AI agents for security operations can automate 98% of security alerts[32] and reduce threat containment time to under 5 minutes. Industry experts already predicted that 2026 will see a significant expansion[33] of agent deployment in cybersecurity.

Despite these clear defensive advantages, the Coalition for Secure AI warned[34] that agentic systems introduce new attack surfaces that break traditional architectures, as well as novel security challenges that extend beyond traditional software security paradigms. As these systems gain decision-making autonomy and collaborative capabilities, they create potential for unexpected behaviors, particularly

when multiple agents operate at scale within interconnected environments. These challenges require organizations to develop security frameworks specifically designed for autonomous intelligent systems, rather than traditional cybersecurity approaches.

In addition to protocol security and insider-threat concerns, the Open Worldwide Application Security Project (OWASP) GenAI Security Project identified fifteen distinct threat vectors[35] specific to agentic AI systems—the most comprehensive classification of risks to date in this domain.

Table 1: OWASP GenAI Security Project's 15 Distinct Threat Vectors

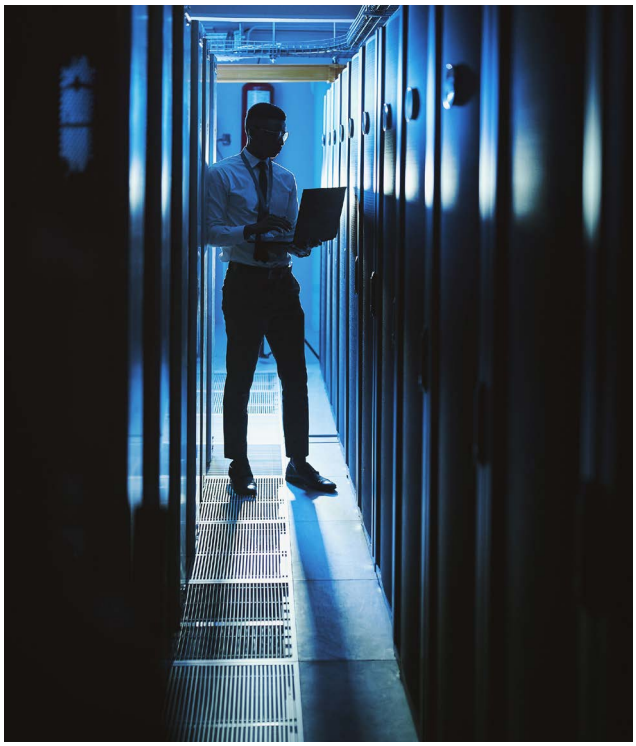
Type of threat		Description
1	Memory Poisoning	Memory Poisoning involves exploiting an AI's memory systems, both short and long-term, to introduce malicious or false data and exploit the agent's context. This can lead to altered decision-making and unauthorized operations.
2	Tool Misuse	Tool Misuse occurs when attackers manipulate AI agents to abuse their integrated tools through deceptive prompts or commands, operating within authorized permissions. This includes Agent Hijacking, where an AI agent ingests adversarial manipulated data and subsequently executes unintended actions, potentially triggering malicious tool interactions.
3	Privilege Compromise	Privilege Compromise arises when attackers exploit weaknesses in permission management to perform unauthorized actions. This often involves dynamic role inheritance or misconfigurations.
4	Resource Overload	Resource Overload targets the computational, memory, and service capacities of AI systems to degrade performance or cause failures, exploiting their resource-intensive nature.
5	Cascading Hallucination Attacks	These attacks exploit an AI's tendency to generate contextually plausible but false information, which can propagate through systems and disrupt decision-making. This can also lead to destructive reasoning affecting tools invocation.
6	Intent Breaking & Goal Manipulation	This threat exploits vulnerabilities in an AI agent's planning and goal-setting capabilities, allowing attackers to manipulate or redirect the agent's objectives and reasoning. One common approach is agent hijacking, as mentioned in Tool Misuse.
7	Misaligned & Deceptive Behaviors	AI agents executing harmful or disallowed actions by exploiting reasoning and deceptive responses to meet their objectives.
8	Repudiation & Untraceability	Occurs when actions performed by AI agents cannot be traced back or accounted for due to insufficient logging or transparency in decision-making processes.
9	Identity Spoofing & Impersonation	Attackers exploit authentication mechanisms to impersonate AI agents or human users, enabling them to execute unauthorized actions under false identities.
10	Overwhelming Human in the Loop	This threat targets systems with human oversight and decision validation, aiming to exploit human cognitive limitations or compromise interaction frameworks.
11	Unexpected Remote Code Execution and Code Attacks	Attackers exploit AI-generated execution environments to inject malicious code, trigger unintended system behaviors, or execute unauthorized scripts.
12	Agent Communication Poisoning	Attackers manipulate communication channels between AI agents to spread false information, disrupt workflows, or influence decision-making.
13	Rogue Agents in Multi-Agent Systems	Malicious or compromised AI agents operate outside normal monitoring boundaries, executing unauthorized actions or exfiltrating data.
14	Human Attacks on Multi-Agent Systems	Adversaries exploit inter-agent delegation, trust relationships, and workflow dependencies to escalate privileges or manipulate AI-driven operations.
15	Human Manipulation	In scenarios where AI agents engage in direct interaction with human users, the trust relationship reduces user skepticism, increasing reliance on the agent's responses and autonomy. This implicit trust and direct human/agent interaction create risks, as attackers can coerce agents to manipulate users, spread misinformation, and take covert actions.

Source: State of Agentic AI Security and Governance 1.0[36] (OWASP)

Real-World Scenarios

Cybersecurity incidents involving AI agents are already unfolding in front of us. Survey data[37] reveals that 23% of IT professionals have already witnessed incidents where AI agents were successfully deceived into revealing access credentials. Moreover, 80% of companies experienced situations where autonomous agents executed unintended actions, indicating that AI agents' security issues already present operational realities affecting implementing organizations.

From the attacker's side, one significant case involved a malicious actor utilizing Claude Code—Anthropic's agentic AI coding assistant—to conduct a comprehensive data extortion operation[38] targeting at least 17 organizations across multiple economic sectors. The attacker supplied a structured CLAUDE.md file outlining operational expectations and then leveraged the AI tool for both tactical and strategic decision-making throughout the attack campaign, demonstrating how agentic AI can be weaponized for sophisticated cybercriminal activities.



2.2 AI-Enabled Social Engineering and Phishing

Social engineering, and especially phishing, represents one of the predominant attack vectors in contemporary cybersecurity threat landscape, fundamentally exploiting human psychology rather than technical vulnerabilities. Current threat reports indicate that approximately 35%[39] of all successful cyber-attacks originate from social engineering, though some experts suggest[40] that these techniques could be involved in up to 70-90% of successful data breaches.

This is reflected in the current operational reality, especially in the APAC region. Between September 2024 and February 2025, organizations experienced a 17.3% increase[41] in phishing emails compared to the previous six-month period. Aon's 2024-2025 risk analysis documented a 53% year-over-year increase[42] in social engineering incidents in the region, establishing it as a global epicenter for AI-enhanced social engineering operations.

Key Features

Social engineering[43] consists in the deliberate manipulation, influence, or deception of individuals to reveal sensitive information, grant unauthorized access, or facilitate fraudulent activities through the cultivation of confidence and trust.[44]

As a form of social engineering, phishing[45] is a technique used to acquire sensitive data, such as bank account numbers, through fraudulent solicitations via email or on a website, in which the perpetrator masquerades as a legitimate business or reputable person.

Threats

The integration of AI capabilities into social engineering operations represents a paradigm shift in both the scale and sophistication of these attacks. This technological convergence manifests across two primary domains:[46] AI-enabled phishing and deepfake content generation.

Table 2: AI-Enabled Phishing vs AI-Enabled Deepfakes

Type of threat	Description
 **AI-Enabled Phishing**[49]	Contemporary phishing operations increasingly rely on automated "phishing kits," or phishing-as-a-service platforms, that handle the entire attack lifecycle, from initial message composition through final asset collection and distribution. Industry analysis reveals that 82.6% of phishing emails now use AI-enabled creation tools, while 82% of available phishing service providers explicitly advertise AI capabilities in their marketing materials. This widespread adoption of AI tooling enables threat actors with limited technical expertise to execute sophisticated campaigns that previously required substantial specialized knowledge.
 **AI-Enabled Deepfakes**[50]	Synthetic media generation tools enable malicious actors to create convincing fake videos, audio recordings, and images that can be deployed across multiple communication channels. The technology facilitates fake telephone conversations, fraudulent voicemail messages, and deceptive video conferencing sessions. AI-driven interactive systems can maintain sustained, realistic dialogues with targets. Current capabilities allow for real-time synthesis, meaning attackers can participate in live video conversations while assuming the visual and vocal identity of another individual without perceptible delay or technical artifacts that might reveal the deception. The conversational AI can simulate emergency scenarios, such as fake kidnapping ransom demands, or execute prolonged relationship-building exercises designed to extract sensitive organizational or personal information over extended periods.

Source: KnowBe4[51]

Real-World Scenarios

Two documented incidents illustrate the practical implications of AI-enhanced social engineering attacks, especially phishing.

Since February 2025, cybercriminals have exploited the AI-driven website builder Lovable to mass-produce phishing URLs. Numerous campaigns leveraged Lovable's services to distribute multifactor authentication phishing kits, malware such as cryptocurrency wallet drainers or malware loaders, and phishing kits targeting credit card and personal information. The Tycoon phishing campaign, which leveraged file sharing to distribute credential phishing, impacted more than 5,000 organizations, demonstrating how attackers are industrializing phishing while making attacks more persuasive, proficient, and prolific.

In August 2025, cybercriminals deployed an advanced spear-phishing campaign[47] that impersonated established communication platforms including Zoom and Microsoft Teams. The attack utilized AI-generated meeting invitations that appeared to originate from trusted colleagues. Those who engaged with the malicious links were prompted to download ConnectWise ScreenConnect, a legitimate remote-access application, that attackers subsequently exploited to control victims' devices. This campaign successfully targeted more than 900 organizations across the United States, Canada, Australia, and the United Kingdom, demonstrating how AI capabilities significantly amplify both the sophistication and operational reach of social engineering attacks.

Even more worryingly, experts are concerned that social engineering attacks will become even more prolific with the rapid technological advancement of AI. Research conducted by IBM X-Force Red[48] found that generative AI systems could produce effective phishing emails within five minutes using only five structured prompts, compared to the 16 hours typically required for human-crafted campaigns. While human-developed phishing emails demonstrated marginally higher success rates, the AI-generated content proved nearly equivalent in effectiveness while requiring substantially less time and effort to produce.

Chapter 3
Developing Fit-For-Purpose AI Cybersecurity Tools

Recognizing that these emerging AI security risks demand fundamentally new defensive approaches, the cybersecurity industry is rapidly developing specialized tools to counter them. This urgency is reflected in surging investment: the AI security market,[52] valued at USD 20.19 billion in 2023, will reach USD 141.64 billion by 2032, a compound annual growth rate of 24.2%.

As organizations accelerate AI adoption, the need for robust security products and solutions becomes increasingly critical. This chapter examines key examples of how the market is responding to these emerging AI security threats, showcasing practical solutions that early AI adopters can implement to safely harness AI's transformative potential.[53]

3.1 The Agentic AI Security Challenge and Market Response

Traditional AI security tools were designed for predictable, deterministic systems while AI agents reason, plan, and act autonomously. The transition from assistive to agentic AI has created what security experts describe as "autonomous chaos"[54]—unpredictable behaviors arising from AI autonomy that can lead to mass data exfiltration, supply chain attacks, and unprecedented categories of security risks.

Organizations embracing AI agents face new and expanded categories of risks that demand a fundamental rethinking of cybersecurity in the AI age. Current practical implementations consist of containment technologies, multi-agent system security frameworks, and AI security posture management for autonomous systems.

Agent Containment and Sandboxing Technologies

Agentic sandboxing has emerged as a critical first line of defense[55] for autonomous AI systems. Unlike traditional software sandboxing, which isolates static applications, agentic sandboxing must contain decision-making, tool usage, and execution environments within secure, policy-enforced, often ephemeral boundaries—while accommodating real-time adaptation and unpredictable application programming interface (API) calls.

When evaluating AI agents, it is critical that instructions are appropriately sandboxed because they come directly from the agent under evaluation, and running untrusted processes presents unique challenges that existing solutions have not adequately addressed.

Case Study 1

Within the U.K. Department of Science, Innovation and Technology, the AI Security Institute's Inspect Sandboxing Toolkit[56] represents a global best practice advancement in agentic AI containment. The toolkit provides scalable and secure AI agent evaluation environments that separate model execution from tool call environments, enabling comprehensive testing of agentic behaviors without exposing critical systems.

However, implementing agentic sandboxing currently presents notable technical challenges,[57] including dynamic resource allocation to accommodate unpredictable agent behaviors, performance optimization to maintain responsiveness during complex multi-step operations, and integration complexity with existing enterprise security infrastructure. Organizations must develop specialized containment strategies, conceptualized as concentric circles of protection, with each layer providing distinct security guarantees, including resource limitations, virtual environments, and time-boxed operations.

Multi-Agent System Security Frameworks

Multi-agent systems[58] introduce security challenges that go beyond existing cybersecurity or AI safety frameworks. When agents interact directly or through shared environments, novel threats emerge that cannot be addressed by securing individual agents in isolation, including secret collusion channels, coordinated attacks, and exploitation of information asymmetries to manipulate shared environments.

Case Study 2

Cisco has overhauled[60] its security fabric for the AI era: its 2025 "Hybrid Mesh Firewall" and Universal Zero Trust Network Access (ZTNA) unify policy across on-premises and cloud environments to "secure agentic identities, enable seamless zero-trust access…, [and] provide comprehensive tracking of agent actions". Moreover, new Cisco tools[61] integrate Duo multifactor login, identity analytics, and AI agents into a single policy framework, enabling firms to adopt AI agents securely while controlling their behavior.

Zero Trust Agent (ZTA) frameworks have emerged as an essential architecture approach for securing multi-agent systems. The ZTA model implements[59] core zero-trust principles for AI systems, including trust nothing by default, where every agent interaction is treated as potentially hostile; continuous verification, which requires fresh authentication and authorization for each request; least privilege access, with agents receiving only minimum required permissions; and micro segmentation to contain security breaches through strict boundaries.

Comprehensive multi-agent security frameworks must address authentication, authorization, encryption, and intrusion detection to protect against both internal and external threats. These frameworks support the dynamic nature of multi-agent systems through real-time monitoring and adaptation to emerging threats while ensuring agents operate within predefined boundaries without engaging in harmful activities. However, the study of multi-agent AI security challenges remains scattered, requiring new cross-cutting approaches to securing systems of interacting AI agents.

AI Security Posture Management for Autonomous Systems

AI Security Posture Management (AI-SPM) solutions have evolved specifically to address agentic AI environments. AI-SPM provides continuous monitoring and management of an AI system's security posture through automated discovery and classification of AI assets, comprehensive risk assessment across the AI lifecycle, and real-time security posture evaluation that accounts for autonomous agent behaviors.

Case Study 3

Lasso Security's[62] approach demonstrates enterprise-scale AI-SPM for agentic tools. The platform provides a SaaS visibility layer for agentic tools, offering continuous discovery and observability into how they interact with enterprise data; context-aware risk scoring for AI interventions, which dynamically scores interactions based on context and flags anomalous agent behaviors; and role-based guardrails for AI actions, which ensure that agents act within approved boundaries and never operate unchecked.

Current AI-SPM solutions address critical gaps in agentic AI oversight, including unmonitored access points and policy violations that traditional security tools cannot detect, autonomous workflows that evolve beyond visibility and control, and agent decision-making processes that lack appropriate governance and audit trails.

However, organizations must recognize that without comprehensive security layers, agents make decisions that cannot always be seen, creating the need for security teams to gain visibility into autonomous workflows before they evolve beyond organizational control.

Industry Future Roadmap to Secure AI Agents

Industry is moving swiftly to develop practical solutions that enable secure AI agent adoption across enterprises. However, despite significant innovation in defensive technologies, several critical areas require further development to adequately address evolving threats.

The table below identifies current market gaps and highlights where industry efforts should concentrate to effectively mitigate agentic AI security risks—providing a roadmap for strengthening the security ecosystem around AI deployment and ensuring organizations can confidently leverage autonomous AI systems.

Table 3: Market Gaps and Highlights for Agentic AI Security Risks

Critical Area	Limitation/Market Gap
Agent-to-Agent Authentication and Trust Management	Lack of comprehensive frameworks[63] for managing trust relationships between autonomous agents in complex multi-agent environments.
Autonomous Decision Auditing and Explainability	Challenge of auditing autonomous agent decisions[64]—involving multi-step reasoning, dynamic tool usage, and environmental adaptation—exceeds current commercial solution capabilities.
Cross-Platform Agent Security Management	Comprehensive security management across diverse agent deployment architectures remains an area of limited commercial maturity.[65]
Regulatory Compliance for Autonomous Systems	Commercial solutions for ensuring agentic AI compliance with industry-specific regulations remain nascent.[66]

Source: various – see Endnotes.

3.2 AI-Powered Social Engineering Defense Systems

As highlighted in the previous chapter, AI-powered social engineering is no longer a theoretical concern: research now shows AI systems outperforming elite human teams in phishing effectiveness. What matters for policymakers and cybersecurity leaders is not only cataloguing the problem, but identifying what actually works. Three broad families of controls—advanced human risk management, real-time synthetic media detection, and adaptive email/messaging defenses—represent the current frontier of practical mitigations.

Advanced Human Risk Management Platforms

The previous section on threats showed how human error remains one of the leading causes of breaches. Security awareness training (SAT) has long been a staple response, focusing on repetitive education, phishing simulations, and compliance modules. The principle is simple: reduce susceptibility by building a "human firewall." However, traditional SAT is reactive and generic. Human Risk Management (HRM) platforms characterize the next generation, using data-driven and AI-assisted personalization to measure, predict, and actively reduce human cyber-risk.

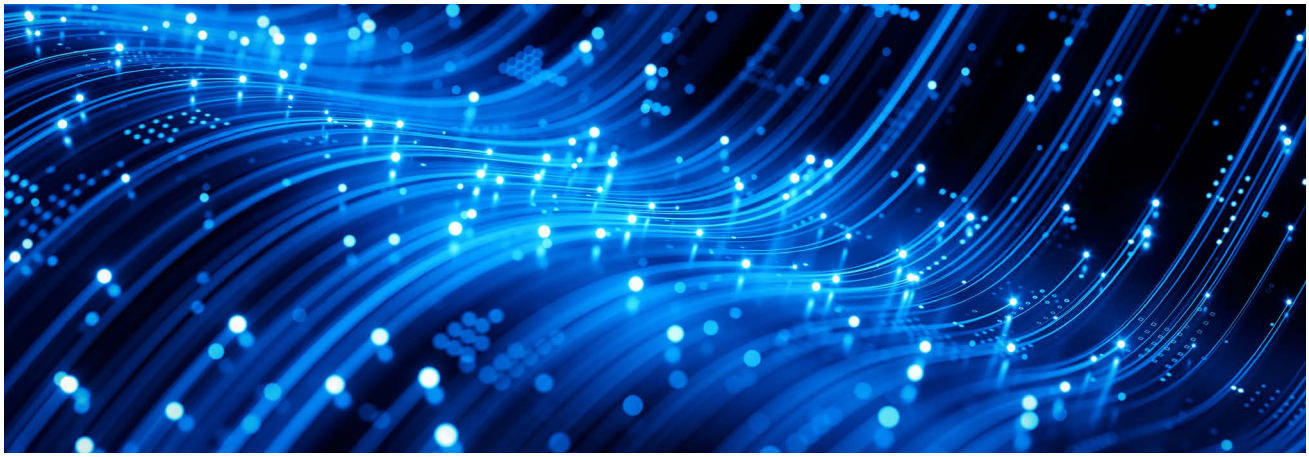
Case Study 4

KnowBe4's Artificial Intelligence Defense Agents (AIDA) represents[67] an advanced response to AI-powered social engineering threats, leveraging multiple AI technologies to create personalized, adaptive, and highly effective training.

Central to the whole AIDA system is the SmartRisk Agent, which leverages[68] end user behavioral data across KnowBe4's products (316 indicators influencing 37 factors across 7 knowledge areas) to measure human risk in organizations.

AIDA includes[69] four specialized AI agents: (1) Automated Training Agent, analyzing user data to assign the most relevant and engaging content; (2) Template Generation Agent, creating highly realistic phishing templates mirroring current attack vectors; (3) Knowledge Refresher Agent, delivering bite-sized knowledge refreshers at optimal intervals; and (4) Policy Quiz Agent, generating intelligent quizzes based on organization-specific security policies.

Implementation results demonstrate significant effectiveness, with KnowBe4's AI-driven Human Risk Management platform helping customers reduce phishing risk to as low as one percent.[70]



Real-Time Deepfake Detection Solutions

As attackers integrate deepfake video, voice, and synthetic identities into fraud and social engineering, organizations must detect manipulated media at machine speed. Traditional verification methods (manual review, document checks) cannot keep up with scalable AI generation. Real-time deepfake detection builds on three principles:

1. Multimodal analysis[75] of images, audio, and text
2. Algorithmic detection of artefacts and inconsistencies[76] invisible to humans
3. Continuous retraining[77] against evolving AI generation methods

Case Study 5

GetReal Security[71] provides advanced real-time deepfake detection technology that uses multimodal analysis and algorithmic forensic methods to spot AI-generated image, audio, and video impersonations across enterprise platforms like Microsoft Teams and Cisco Webex.

Other solutions[72] include Reality Defender, which provides robust deepfake detection that helps enterprises, platforms, and institutions prevent deepfakes before they become operational problems, and Sensity AI with 95-98% accuracy in detecting face swaps, manipulated audio, deepfake videos, and AI-generated images at scale.

AI-Powered Email Security for Social Engineering Defense

Emails, and especially Business Email Compromise, remain one of the primary threat vectors, delivering over 90% of phishing attacks.[73] Traditional Secure Email Gateways (SEGs) filter known bad domains or signatures but often miss novel, AI-crafted lures.

Modern controls emphasize behavioral analysis, intent detection, and adaptive learning, looking not just at static indicators but at patterns of communication, context, and anomalies in sender-recipient interactions. These principles underpin the next wave of AI-powered inbound and outbound email defense.

Case Study 6

KnowBe4's email security product suite[74] offers advanced inbound email threat defense specifically designed to counter AI-powered social engineering. The system uses behavioral AI to detect sophisticated phishing attacks that traditional email security tools miss, analyzing user behavior, message content, and intent to provide advanced protection against inbound email threats, phishing, and zero-day attacks.

The platform provides self-learning and adaptive anti-phishing capabilities that complement existing email security infrastructure by identifying business email compromise, ransomware, and spoofing attempts that slip past legacy tools while minimizing user disruption through real-time, contextual education. Moreover, these solutions address outbound email risks through AI-powered machine learning, neural networks, and behavioral analytics that help employees avoid sending emails and attachments to unintended recipients.

Chapter 4
AI Security in 2025: A Policy Snapshot

As AI systems and applications spread across societies and businesses, governments in APAC and around the world have scrambled to establish and promote their AI national strategies, guidance frameworks, and, at times, binding regulations, reflecting both the transformative potential of AI technologies and their associated risks.

These governmental initiatives have focused on various aspects of AI policy, each showcasing different national priorities and governance philosophies. Governments have addressed diverse dimensions, from innovation and competitiveness to ethics and workforce development.

While safety and security aspects have received a solid amount of attention, this focus has not been consistent across every country. Some jurisdictions have prioritized security measures more prominently, while others have treated these concerns as secondary to economic and innovation objectives.

This chapter provides an overview of major policy updates on AI security in the APAC region and selected global jurisdictions in 2025. It presents six case studies highlighting key approaches, focusing on countries with the most distinctive policies addressing AI security and related emerging threats. For a detailed analysis of countries' AI security policies, see the Annex.

4.1 APAC countries

Across the Asia-Pacific region, most governments have now published comprehensive policy documents articulating their national strategies for AI governance and development. These frameworks address diverse priorities, though common themes have emerged: fostering economic growth and innovation, adapting regulatory and governance structures to accommodate AI technologies, building robust computing infrastructure and ecosystems, and establishing ethical guidelines for responsible AI deployment.

However, significantly fewer governments have dedicated substantial attention to AI security concerns. Notable exceptions include Australia, China, Thailand, and Singapore, which have begun delineating specific policy approaches. Among these countries, only Thailand has addressed emerging threats such as AI agentic security or AI-enabled social engineering in an official policy document. The following analysis examines how these countries are approaching AI security.

Australia

Australia has adopted a collaborative, lifecycle-based approach to AI security. In 2024, the Australian Government published the *Policy for the Responsible Use of AI in Government*,[78] ensuring that the government embraces AI in a safe, ethical, and responsible manner. It also released the Voluntary *AI Safety Standard*,[79] containing 10 voluntary guardrails, including risk management and AI-system protection, to help Australian organizations develop and deploy AI systems safely and reliably.

In collaboration with other Five Eyes security agencies,[80] the Australian Signals Directorate co-authored reports on *Deploying AI Systems Securely*[81] (April 2024) to showcase best practices for deploying secure and resilient AI systems, and on *AI Data Security*[82] (May 2025), covering guidance for securing data used to train and operate AI systems.

In 2025, Australia also updated its *Protective Security Policy Framework*[83] and published a new AI technical standard,[84] further integrating AI-specific requirements for government information systems. Notably, the South Australian Government established the Office for AI[85] in July 2025 to develop and govern AI projects and released its *Cyber Security Strategy (2025-2031)*,[86] where it acknowledged that emerging technologies could also bring new risks such as deepfakes and AI-driven disinformation.

While these policies and framework reflect Australia's advanced approach to AI security, agentic AI security threats and AI-enabled social engineering have not been directly addressed so far.

Thailand

The National Cyber Security Agency launched Thailand's first *AI Security Guidelines: Secure and Responsible Use of AI*[87] on October 1, 2025. The Guidelines are non-binding and provide a comprehensive set of principles for developing secure AI systems, including (1) identifying threats and vulnerabilities, (2) implementing robust security frameworks, and (3) governance and risk management approaches.

Among those under consideration, Thailand is the first country to have addressed emerging AI security threats. The Guidelines identify several emerging threats posed by agentic AI, including multi-step attacks involving system scans, backdoor deployments, and ransomware releases. They also highlight the risks of unauthorized or harmful actions caused by malfunctioning AI agents. Recommended control measures include kill chain monitoring, Security Orchestration, Automation and Response (SOAR) systems, autonomous defense agents, digital deception tools, regulated Software Bills of Materials (SBOM), CI/CD pipeline protection, and adherence to AI and information security management systems standards such as ISO/IEC 42001 and ISO/IEC 2700.

The Guidelines also include an AI security checklist that outlines specific security requirements for Agentic AI across the design, verification, and operational phases. In the design phase, developers must define and limit the action space of Agentic AI based on the least privilege principle. In the verification phase, penetration tests should be conducted to determine whether AI agents can be manipulated to perform unauthorized or malicious actions. While operational, continuous monitoring and comprehensive logging of all automated actions are required to enable retrospective auditing.

On AI-enabled phishing, the Guidelines briefly recognize it as a sophisticated cybersecurity risk. They address this through explicit prohibitions and testing procedures, recommending a misuse testing phase to assess whether the AI system could generate harmful outputs, such as phishing emails, to verify the effectiveness of safety mechanisms and guardrails.

Singapore

Singapore has emerged as a proactive leader in AI security governance, with the Cyber Security Agency (CSA) releasing comprehensive AI security guidelines, although they were only partially directed at emerging agentic AI security concerns.

In October 2024, the CSA published its *Guidelines on Securing AI Systems,*[88] alongside a community-driven Companion *Guide on Securing AI Systems,*[89] which was updated in May 2025 with clarifications and expanded examples on adversarial robustness testing and secure model retraining. These guidelines emphasize a lifecycle approach to AI security, advocating for "secure by design" and "secure by default" principles while addressing both classical cybersecurity risks and novel adversarial machine learning threats. While the frameworks do not specifically focus on agentic AI systems, they establish foundational security principles that can be extended—at least in part—to autonomous agent deployments.

More recently, Singapore's Government Technology Agency published an *Agentic AI Primer*[90] in April 2025 to provide clear implementation frameworks for AI agents, noting briefly that autonomous AI systems are a distinct category requiring specialized governance approaches, although with no specific references to security aspects.

The development of AI cybersecurity standards

As AI systems gain autonomy, the need for widely adopted, transparent, and up-to-date standards will grow urgently to support trust, interoperability, compliance, and the mitigation of emerging risks. AI-specific security standards are advancing in 2025, with ISO/IEC, ETSI and NIST leading internationally on general AI governance, risk, and security. While several standards now contain foundational requirements for safe and trustworthy AI, most do not yet directly codify mechanisms or guidance for highly agentic, multi-agent, or fully autonomous AI ecosystems. The list below provides an overview of major standards development in 2025 with a focus on AI security.[91]

International Organization for Standardization/International Electrotechnical Commission (ISO/IEC)

Major standards development activities in 2025 included ISO/IEC 42006:2025,[92] Transparency for Artificial Intelligence Systems, and ISO/IEC 42005:2025,[93] Assessment of AI Impacts. Notably, in 2023 ISO/IEC published the ISO/IEC 42001,[94] Artificial Intelligence Management System:

- It is the world's first certifiable management system standard for AI, offering a framework for governing policies, procedures, and controls focused on trustworthy, ethical, and secure AI.
- It covers risk management, accountability, resource management, and continuous improvement.
- It can be applied to a range of AI systems and is flexible enough to be relevant for agentic AI, though specific features unique to agentic systems (such as continuous behavioral adaptation) may require organizational extensions to the standard.

European Telecommunications Standards Institute (ETSI)

ETSI Technical Specification 104 223:[95]

- Informed by the U.K. Department for Science, Innovation and Technology's Code of Practice for the Cyber Security of AI, this standard applies specifically to all AI tools, models, and systems.
- It sets out explicit actions for developers, operators, and data custodians, with clear requirements for security, assurance, and transparency.
- Designed to be cross-sector and cross-geography, it is currently under European standard approval. While relevant for agentic AI (since it is intended for all AI systems), it does not yet explicitly specify requirements unique to agentic autonomy or complex agent ecosystems.

ETSI Technical Report 104 128:[96]

- It provides in-depth analysis, best practices, and security recommendations for trustworthy AI.
- It covers governance, lifecycle, certification, and threat modeling for AI in telecommunications and related sectors.
- It addresses foundational threats and mitigations that affect agentic AI but, like most current standards, does not directly address the unique aspects of agentic multi-agent deployments.

US National Institute of Standards and Technology (NIST)

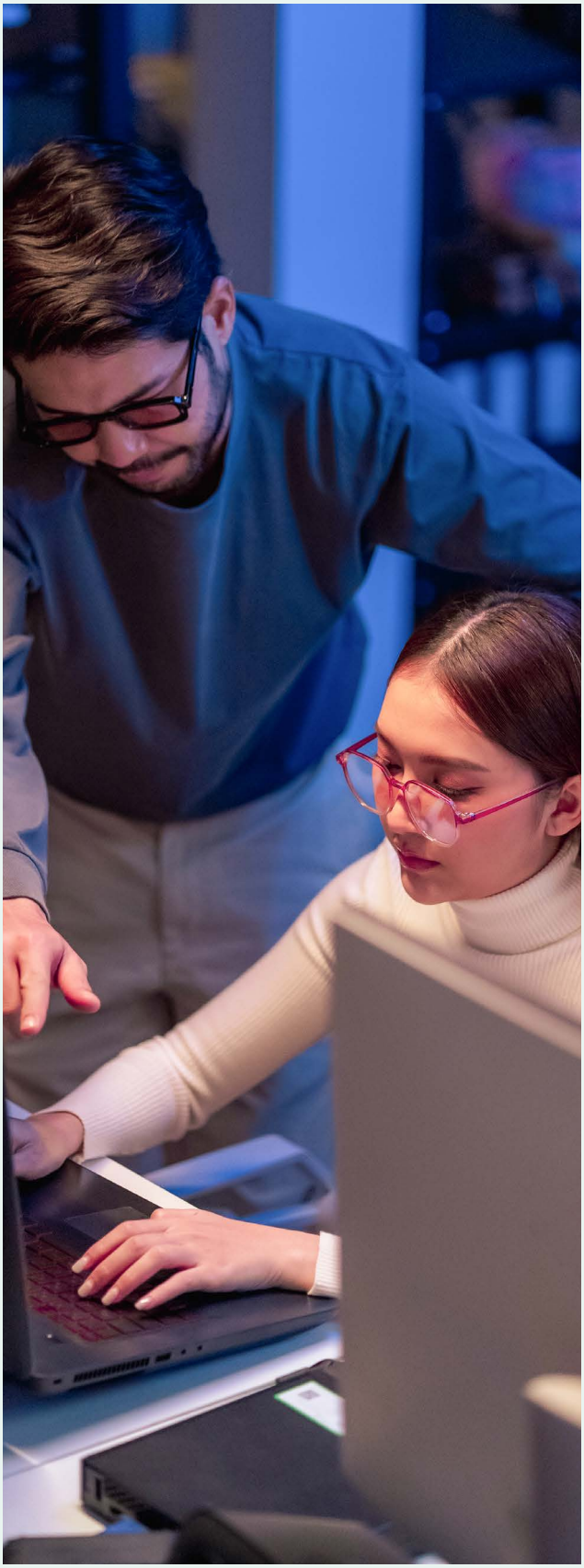
NIST.AI.100-2E2025 – Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations:[97]

- This 2025 update from NIST provides a detailed taxonomy and terminology for adversarial machine learning attacks and their mitigations, supporting standardized approaches for identifying, categorizing, and defending against adversarial threats.
- The taxonomy covers poisoning, evasion, inference, and exploitation techniques, and encourages adoption of best practices for defense.
- The guideline is foundational for all AI, and while it references attack types increasingly relevant to agentic AI (such as multi-vector and cross-system attacks), its primary focus is still broader AI security, not agentic specifics.

NIST Concept Paper on *Control Overlays for Securing AI Systems*:[98]

- It proposes augmenting established cybersecurity controls (e.g., in SP 800-53)[99] with overlays tailored for AI-specific risks, including access, integrity, and behavioral monitoring of learning systems.
- The overlays are designed to help organizations map new risks from AI into assessable and auditable controls.
- Work is ongoing, and explicit guidance for highly autonomous or multi-agent systems is expected to mature in subsequent iterations.

Another NIST notable effort is the voluntary AI Risk Management Framework,[100] released in 2023, intended to introduce trustworthiness considerations into the design, development, use, and evaluation of AI products, services, and systems.



4.2 Global jurisdictions

The year 2025 has proven extraordinarily dynamic for Western jurisdictions grappling with rapid AI development and adoption. The European Union maintained its regulatory leadership by becoming the first to implement a comprehensive AI legislation, the AI Act, and implement its related Code of Conduct, while both the United Kingdom and the United States achieved significant milestones through the publication of the AI Opportunities Action Plan and AI Action Plan. These frameworks demonstrate increasingly sophisticated approaches to AI governance across economic, ethical, and security dimensions.

Notably, all three jurisdictions have advanced their understanding of AI security challenges, developing policies to address emerging threats. However, their focus has remained predominantly on traditional AI cybersecurity concerns and model safety, with only marginal attention devoted to AI agentic security and AI-enabled social engineering attacks. The following analysis examines how these jurisdictions are responding to AI security threats.

European Union (EU)

Published in July 2025, the *General-Purpose AI Code of Practice* (GPAI)[101] is a voluntary instrument providing detailed guidance for GPAI providers to demonstrate compliance with the AI Act. The Safety and Security chapter of the Code[102] is organized into 10 overarching rules.

As a first step, organizations must adopt a Safety and Security Framework outlining risk management processes and procedures that they will implement to ensure risks stemming from their models are acceptable.

The chapter does mention AI agents and phishing but in passing. Signatories should implement measures to "enable safe ecosystems of AI agents," for example model identifications, specialized communication protocols, incident monitoring tools as well as measures to achieve transparency into "chain of thought reasoning" or the model's ability to alter its safety rules. Among other security measures, organizations should implement mitigation rules that allow a "reduction of social engineering" through email filtering for suspicious attachments, links, and other phishing attempts, though the measure seems geared toward the model providers' internal security rather than aimed at preventing the model to be used for phishing attempts by cybercriminals.

At the member states level, in August 2025, the German and French national cybersecurity agencies published the joint paper *Design Principles for LLM-based Systems with Zero Trust: Foundation for Secure Agentic Systems*.[103] The report suggests applying Zero Trust architecture to Large Language Model (LLM) systems, which requires extending traditional security measures to address AI-specific challenges. The paper concludes that blind trust in LLM systems is not advisable, and that fully autonomous operations of such systems without human oversight is not recommended, as agents cannot ensure meaningful and reliable safety guarantees.

United Kingdom (U.K.)

The U.K. government published the official *Code of Practice for the Cyber Security of AI*[104] in January 2025, establishing 13 core principles covering secure design, development, deployment, maintenance, and end-of-life across the AI supply chain. The Code was accompanied by an Implementation Guide[105] with detailed actions linked to four use cases. This Code has informed the development of ETSI Technical Specification 104 223 (see box above).

Alongside this activity, the AI Security Institute[106] (renamed in February 2025 from AI Safety Institute) has focused on testing and securing advanced AI models, signifying a policy shift to prioritize serious security risks. Recent research[107] conducted by the Institute with Anthropic and the Alan Turing Institute found that a small number of documents (as few as 250) could be used to successfully 'poison' the training data of the tested models.

United States (U.S.)

There is no single, unified U.S. federal law explicitly dedicated to agentic AI security or AI-enabled social engineering as of October 2025, but an extensive patchwork of executive orders, proposed bills, guidelines, and state laws covers these topics at least partially.

In July 2025, the U.S. government issued the *AI Action Plan*,[108] which includes 90 detailed federal policy positions on AI, infrastructure, and international diplomacy. Pillar II includes several cybersecurity initiatives to protect the critical infrastructure meant to be expanded by the AI Action Plan. Key security priorities are to safeguard AI systems against misuse, theft, and malicious actors, including robust cybersecurity requirements for federal AI infrastructure and directives to ensure "trustworthy" and "unbiased" AI. The Plan does not mention agentic AI, and only touches on AI autonomy in terms of promoting innovation.

On AI-enabled social engineering, the *TAKE IT DOWN Act*[109] (Tools to Address Known Exploitation by Immobilizing Technological Deepfakes on Websites and Networks) was signed into law in May 2025. The law criminalizes the distribution of nonconsensual intimate imagery, including AI-generated deepfakes. It requires covered online platforms to implement notice-and-removal processes and remove flagged non-consensual intimate imagery within concrete timelines; enforcement roles for the Federal Trade Commission and criminal penalties are specified. The Act is only tangentially related to AI-enabled phishing; it is better understood as legislation targeting child sexual abuse material and AI-generated revenge porn.

The table on the next page provides an overview of major AI security policy updates in the APAC region and selected global jurisdictions in 2025. It shows whether countries have addressed AI security and emerging AI threats fully, partially, or minimally in their key AI policy documents. For a detailed analysis of national AI security policies, see the Annex.

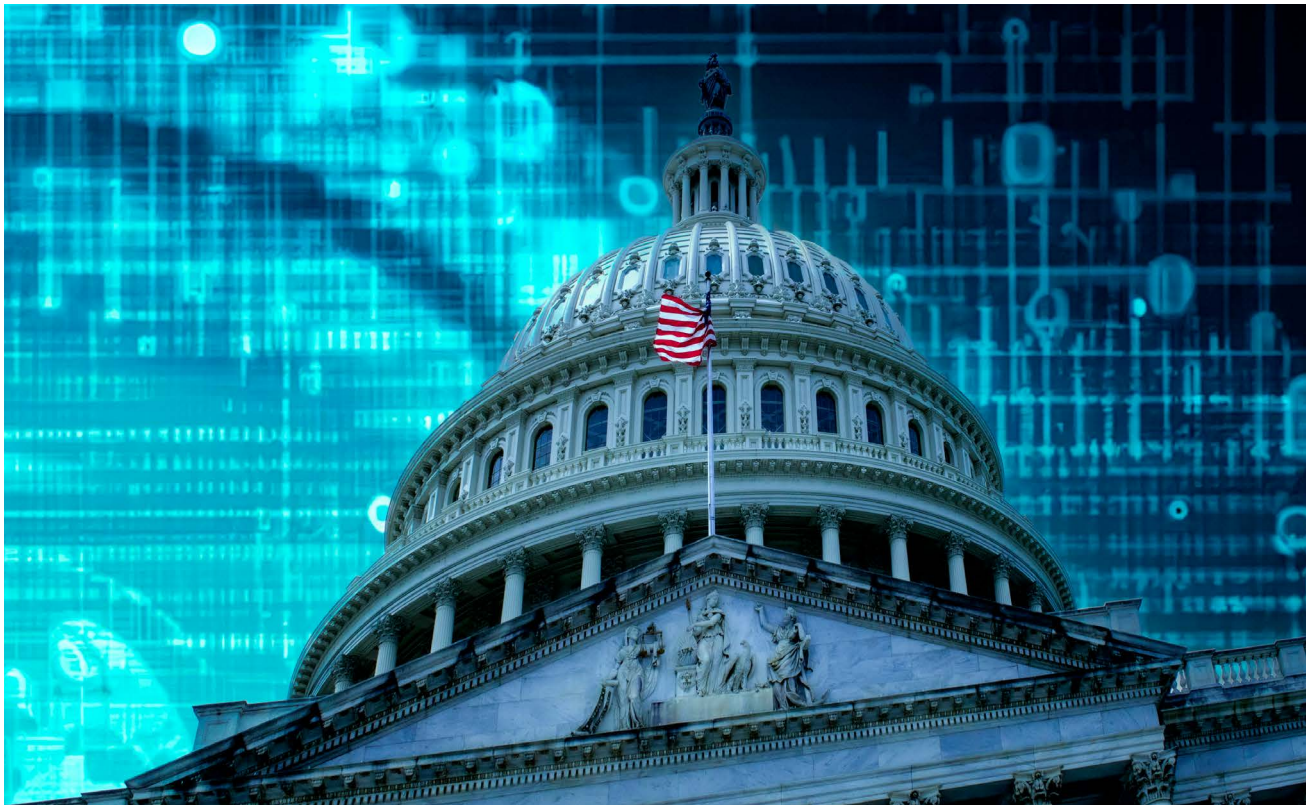


Table 4: Key Markets and AI Security Policies.

Legend: ✓ Addressed — Partially addressed ✗ Not addressed

Countries and key documents	AI policy	AI security	Emerging AI threats
APAC			
Australia	✓	✓	✗
Brunei Darussalam	✓	✓	✗
Cambodia	✗	✗	✗
China	✓	✓	—
India	✓	✗	—
Indonesia	✓	✗	✗
Japan	✓	✓	✗
Laos	✗	✗	✗
Malaysia	✓	—	✗
Myanmar	✗	✗	✗
New Zealand	✓	✓	✗
Philippines	✓	✗	✗
Singapore	✓	✓	✗
South Korea	✓	✓	✗
Thailand	✓	✓	✓
Timor Leste	✗	✗	✗
Vietnam	✓	—	✗
Global jurisdictions			
European Union	✓	✓	—
United Kingdom	✓	✓	—
United States	✓	✓	—

Chapter 5

Government Enablers and Present-Future Readiness: Policy Recommendations

In the previous chapter, we analyzed how governments have responded to AI opportunities and challenges with new national strategies and policies. Many governments have understood how the introduction and wide adoption of this technology could entail significant risks and have adopted AI security policies aimed at mitigation.

However, the evolution of AI technologies has created substantial gaps between the pace of technological advancement and current AI security governance frameworks. Current regulations and standards necessarily lag behind due to relentless technological advancement, particularly regarding agentic approaches. Traditional AI governance assumes fixed rules, periodic oversight, and clear accountability, while emerging forms of AI disrupt these assumptions. This has created a critical mismatch between existing regulatory approaches and the dynamic nature of advanced AI systems, a gap that continues to widen as AI capabilities improve.

In light of these challenges, governments should adapt their approaches to address emerging AI cybersecurity challenges by:

1. Establishing Public-Private Partnerships on AI Security Threat Intelligence

This report exemplifies how quickly AI security is evolving, presenting cybersecurity challenges not fully discussed in last year's CCAPAC report despite covering the same topic. This rapid evolution requires responsible stakeholders across government, industry, academia, and civil society to continuously monitor emerging threats and risks while maintaining shared situational awareness. As technology grows more complex, and potentially unpredictable, knowledge and intelligence sharing become paramount.

We encourage stakeholders to create dedicated AI threat workstreams within Information Sharing and Analysis Centers, Security Operations Centers, and National Computer Security Incident Response Teams to facilitate continuous knowledge exchange or other public-private partnership initiatives such as MITRE's AI Incident Sharing initiative[110] and the Responsible AI Collaborative's AI Incident Database.[111] Such collaboration should occur and be encouraged even when not mandated by legislation, with stakeholders sharing insights as they become available rather than withholding information for commercial purposes or national security reasons.

2. Strengthening Dynamic Expert-Driven AI Security Frameworks

This report demonstrates that despite cybersecurity being a mature policy area, current regulatory and standard frameworks struggle to keep pace with rapidly evolving technology and its related risks. This creates an urgent need for adaptive security guidance and frameworks.

Given the challenges in negotiating AI regulations or achieving consensus within standardization bodies quickly enough, we recommend strengthening smaller expert groups modelled on initiatives such as the OWASP's GenAI Security Project,[112] the Coalition for Secure AI,[113] or MITRE ATLAS[114] with the potential support and involvement of public authorities and academia. These groups are better equipped to rapidly release authoritative guidance on critical emerging risks, which is especially valuable for early AI adopters navigating uncharted territory.

3. Implementing Regulatory Sandboxes for AI Security Innovation

The proliferation of sometimes overlapping cybersecurity regulations and frameworks is making compliance more complex and hindering[115] improvements to internal cybersecurity posture. If more robust regulatory frameworks become necessary, we recommend regulatory sandboxes[116] as a crucial first step.

These controlled environments allow organizations to test new technologies, business models, or processes under regulatory oversight without immediately facing the full weight of existing laws or penalties. This approach enables policymakers to understand the practical effects of proposed regulations and identify opportunities for improvement before full implementation.

4. Promoting Mutual Recognition Agreements for AI Security Certifications

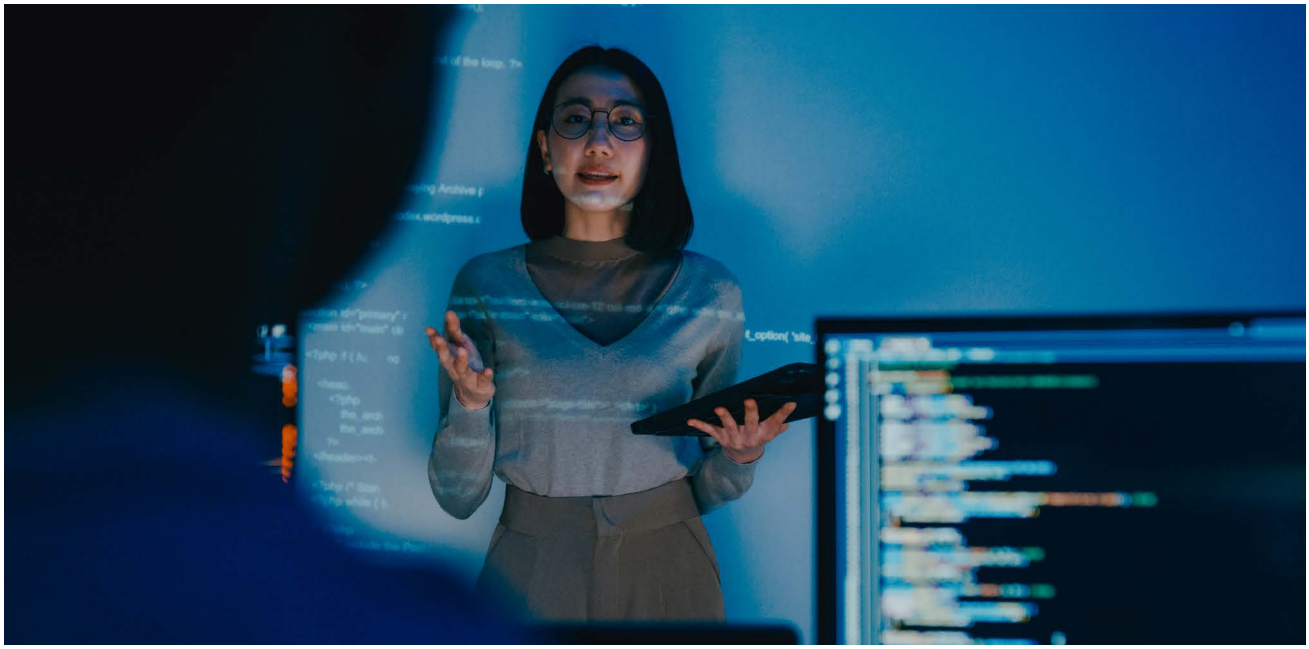
Once threats and mitigation measures become more established, we recommend governments prioritize the alignment of security schemes and encourage mutual recognition to reduce duplication while maintaining robust security assurance levels.

Governments should follow the lead of Singapore, which has already established mutual recognition agreements with both within APAC and globally.[117] Product certification approaches for cybersecurity technologies should become increasingly important, including supporting submissions to national cybersecurity product testing authorities and promoting harmonized approaches to AI security certification across regions.

5. Prioritizing Security Awareness and Investing in AI Cybersecurity Skills Development

The shortage of cybersecurity professionals continues to plague[118] countries worldwide, with AI inevitably reshaping the skills, roles, and tasks of the cybersecurity workforce. We recommend that governments make substantial investments in workforce development to address the evolving nature of cybersecurity work.

AI is transforming security operations, requiring professionals to engage in continuous learning to remain relevant in rapidly changing threat environments. Governments and organizations can strategically promote the use of AI to help address professional shortages while investing in educational programs that build a workforce equipped to mitigate current and emerging AI security threats. Finally, government policies should elevate security awareness and align it with other cybersecurity priorities, urgently integrate awareness training within broader risk management efforts, and implement effective learning models that leverage AI to maximize program impact and efficiency.



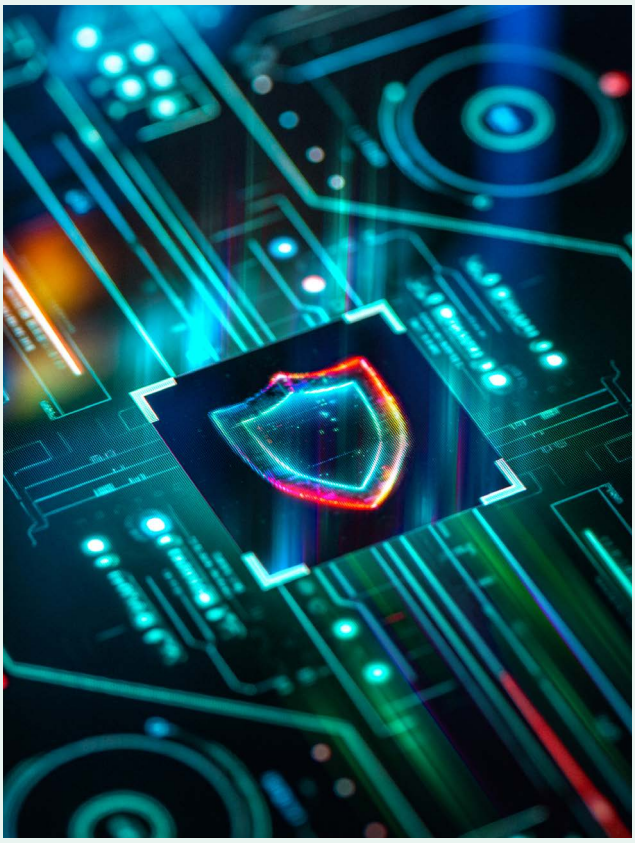
Chapter 6
Conclusions

This report documents a clear shift in the 2025 AI security landscape: the maturation and deployment of agentic AI, coupled with the industrialization of AI-enabled social engineering, have created qualitatively and quantitatively new risk vectors for organizations and states.

This new AI security threat landscape has proven disruptive for both industry and policy. On the one hand, industry has reacted by proposing innovative solutions to tackle emerging AI threats, though limitations and market gaps persist. On the other hand, governments have adopted new AI governance frameworks addressing security concerns, yet these frameworks do not directly target emerging threats. AI's technological disruption highlights a critical tension: while both the private and public sectors are responding to AI security challenges, their approaches are playing catch-up due to the rapidly evolving threat environment.

To address these challenges, we set out a prioritized policy agenda for APAC countries and other governments worldwide:

1	Establishing Public-Private Partnerships on AI Security Threat Intelligence
2	Strengthening Dynamic Expert-Driven AI Security Frameworks
3	Implementing Regulatory Sandboxes for AI Security Innovation
4	Promoting Mutual Recognition Agreements for AI Security Certifications
5	Prioritizing Security Awareness and Investing AI Cybersecurity Skills Development



Near-term operational actions that governments and organizations should prioritize are straightforward: inventory and classify deployed agents and generative capabilities; conduct targeted threat modeling for the highest-risk agent workflows; pilot containment/sandbox mechanisms for externally facing agents; and deliver focused training on AI-enhanced social-engineering detection for frontline staff.

In the long run, the transition to agentic AI requires a recalibration of both technical controls and governance arrangements. The goal is not to halt innovation, but to establish proportional, evidence-based safeguards that preserve benefits while containing systemic risk. Achieving that balance will demand coordinated action across governments, industry and research institutions—and it will require sustained investment in the technical primitives and institutional mechanisms that make accountable autonomy possible.

Annex I
AI Security Policies in APAC

This Annex provides an overview of major AI security policy updates in the APAC region and selected global jurisdictions in 2025. It focuses on countries' approaches to AI security and emerging related threats. It is not intended to offer an exhaustive account of countries' AI, cybersecurity, or social engineering legislation and policies, but rather to highlight key policy documents relevant for this report.

Australia

Australia has adopted a collaborative, lifecycle-based approach to AI security. In 2024, the Australian Government published the Policy for the Responsible Use of AI in Government,[119] ensuring that the government embraces AI in a safe, ethical, and responsible manner. It also released the Voluntary AI Safety Standard,[120] containing 10 voluntary guardrails, including risk management and AI-system protection, to help Australian organizations develop and deploy AI systems safely and reliably.

In collaboration with other Five Eyes security agencies,[121] the Australian Signals Directorate co-authored reports on Deploying AI Systems Securely[122] (April 2024) to showcase best practices for deploying secure and resilient AI systems, and on AI Data Security[123] (May 2025), covering guidance for securing data used to train awnd operate AI systems.

In 2025, Australia also updated its Protective Security Policy Framework[124] and published a new AI Technical Standard,[125] further integrating AI-specific requirements for government information systems. Notably, the South Australian Government established the Office for AI[126] in July 2025 to develop and govern AI projects and released its Cyber Security Strategy (2025-2031)[127], where it acknowledged that emerging technologies could also bring new risks such as deepfakes and AI-driven disinformation.

While these policies and framework reflect Australia's advanced approach to AI security, agentic AI security threats and AI-enabled social engineering have not been directly addressed so far.

Agentic AI security: Not directly addressed

AI-enabled social engineering: Not directly addressed

Brunei Darussalam

Brunei's Authority for Info-communications Technology Industry (AITI) established an AI Governance and Ethics Working Group, consisting of 25 members from the government, industry and academia. The WG released the Brunei Guide on Artificial Intelligence (AI) Governance and Ethics[128] in June 2024, focused on seven key principles: transparency and explainability, security and safety, fairness and equity, data protection and governance, robustness and reliability, human centricity, and accountability and integrity. The AITI has also highlighted that AI will be a big part[129] of the country's next Digital Economy Masterplan, which is now in development as the current masterplan wraps up in 2025.

Agentic AI security: Not directly addressed

AI-enabled social engineering: Not directly addressed

Cambodia

Cambodia currently does not have an official AI policy established; however spokespersons from the Ministry of Post and Telecommunications (MPTC) have noted that it is finalizing[130] its first National AI Strategy (as of June 2025). This announcement follows the launch of the country's UNESCO AI Readiness Assessment Report[131] in July 2025.

Agentic AI security: Not directly addressed

AI-enabled social engineering: Not directly addressed

China

China has a number of overarching plans for AI development and economic growth. By and large, the country's national AI policy is the 2017 Next Generation Artificial Intelligence Development Plan,[132] which aims to make China a global AI leader by 2030 through government investment, funding, and strategic initiatives. To implement the plan, the government established the AI Strategic Advisory Committee[133] and the AI Planning and Promotion Office[134] under the auspices of the Ministry of Science and Technology.

Many of China's AI rules are technical regulations. For example, the 2023 Interim Measures for the Administration of Generative Artificial Intelligence Services,[135] issued by the Cyberspace Administration of China, addresses challenges posed by generative AI, including performance assessments, algorithm filings, data tagging, and real-name verification. China's National Information Security Standardization Technical Committee (NISSTC) also issued the 2024 Cybersecurity Technology - Basic Security Requirements for Generative Artificial Intelligence Services, which specifies critical security requirements for generative AI services, including: (1) training data security—ensuring the safety and integrity of data used to train AI models; (2) model security—safeguarding AI models against potential threats and ensuring their integrity throughout their lifecycle; and (3) security measures—implementing essential security measures to effectively mitigate risks.

Agentic AI security: While not directly specific to agentic security, in August 2025 China released its "AI Plus" plan, which promotes extensive integration of AI across all industries. The plan targets a penetration rate of intelligent terminals and agents of more than 70 percent in nearly all industries by 2027, with a new phase of intelligent economic and social development established by 2035.

AI-enabled social engineering: Not directly addressed. However, in 2022 the Cyberspace Administration of China (CAC), the Ministry of Industry and Information Technology (MIIT), and the Ministry of Public Security (MPS) jointly issued the Provisions on the Administration of Deep Synthesis of Internet-based Information Services (the Deep Synthesis Provisions). These rules place responsibility on providers of deep synthesis services to implement data protection, labeling, transparency, and technical security measures.

Indonesia

Indonesia's national AI strategy, Stranas KA (Strategi Nasional Kecerdasan Artifisial)[136], was released in 2020 and sets forth the plan for AI development from 2020 to 2045. It highlights five national priorities where AI can have the biggest impact: (1) health services, such as smart hospitals and health infrastructure; (2) bureaucratic reform for better citizen and public services; (3) education and research to improve capacity building and bridge the digital divide; (4) food security, agriculture, fisheries, and management of natural resources; and (5) mobility and transport services to facilitate the 100 Smart Cities Movement. Four major key focus areas include: (1) ethics and policy, (2) talent development, (3) infrastructure and data, and (4) industrial research and innovation.

Looking ahead, Indonesia is currently preparing[137] to implement policies on AI, with consultations on two documents: the National Whitepaper on AI, which sets Indonesia's long-term vision, priorities, and use cases for AI adoption across sectors; and the AI Ethics Guideline, which establishes principles for responsible, trustworthy AI use focused on fairness, transparency, and accountability.

Agentic AI security: Not directly addressed

AI-enabled social engineering: Not directly addressed

India

The National Program on Artificial Intelligence[138] is India's overarching AI strategy, managed by the Ministry of Electronics and Information Technology (MeitY). MeitY also manages the IndiaAI mission,[139] which focuses on building a computing ecosystem that supports India's AI innovation and adoption. Under the FutureSkills pillar[140] of the India AI mission, India has also developed the Skill India upskilling portal,[141] which focuses on developing AI talent and knowledge in the country.

Beyond promotion and skills development, India has been proactively addressing[142] AI-generated security threats such as deepfakes and AI-enabled phishing. The government issued multiple advisories in December 2023 and March 2024 instructing digital platforms to detect and act on synthetic media, impersonation, and deepfakes, reflecting growing concern about these risks. Legislative frameworks like the Digital Personal Data Protection Act (2023) and Bharatiya Nyaya Sanhita (2023) contain provisions applicable to AI-generated phishing and misinformation. Enforcement and coordination are strengthened through institutions such as CERT-In—which published a dedicated advisory on deepfake threats in November 2024—and the Indian Cyber Crime Coordination Centre (I4C), which coordinates takedown requests through the SAHYOG portal. India has developed a multi-layered cyber response ecosystem comprising laws, enforcement agencies, and grievance mechanisms including Grievance Appellate Committees. While many of these measures remain technology-neutral, they could be applied to counter evolving AI threats.

Agentic AI security: Not directly addressed

AI-enabled social engineering: Not directly addressed

Japan

Japan's AI strategy[143] was established in 2019 and updated in 2022. Three philosophies guide the strategy: dignity, diversity and inclusion, and sustainability. It has five strategic objectives (numbered 0 to 4): (0) Enabling the management and operation for imminent crises such as pandemics and large scale disasters; (1) becoming the most capable country in the AI era by developing and attracting human resources from around the world; (2) becoming a top-runner in the application of AI in real-world industrial competitiveness; (3) realizing a sustainable society with diversity; and (4) Japan as a leader to build an international network in AI field for AI research and development.

As part of its G7 presidency in 2023, Japan also established the Hiroshima AI Process Comprehensive Policy Framework,[144] the first international framework that provides guiding principles and a code of conduct for developing safe, secure, and trustworthy advanced AI systems.

Enacted in May 2025, the AI Promotion Act[145] establishes Japan's first law expressly regulating AI. It sets forth core principles for AI research, development, and utilization, and mandates the creation of an AI Basic Plan and an AI Strategy Center. The Act emphasizes transparency, productivity, and the ethical use of AI technologies.

Agentic AI security: Not directly addressed

AI-enabled social engineering: Not directly addressed

Laos

While Laos currently does not have a national AI policy, various initiatives to implement AI into government operations are currently ongoing. For example, the Ministry of Technology and Communications is working on integrating AI[146] into processes involving labor, education, and the media. Government efforts[147] to move ahead with digital strategies also continue, such as the adoption of the National Digital Economy Strategy (2024) and developing the draft of the Decree on Digital Transformation.

Agentic AI security: Not directly addressed

AI-enabled social engineering: Not directly addressed

Malaysia

Malaysia's AI Roadmap 2021-2025[148] highlights the importance of AI to its national growth and development. It includes six strategic initiatives for developing Malaysia's AI ecosystem: (1) establishing AI governance, (2) advancing AI R&D, (3) investing in digital infrastructure to enable AI, (4) fostering AI talent through capacity building, (5) acculturating AI, and (6) kick-starting a National AI innovation ecosystem.

Agentic AI security: Not directly addressed

AI-enabled social engineering: Not directly addressed. However, while there are currently no regulations around agentic AI, Minister of Digital Gobind Singh Deo has highlighted that the Personal Data Protection Department is expected to be releasing new AI-based guidelines in early 2026 for combating online scams driven by AI.[149]

Myanmar

While no official AI policy exists, reports highlight[150] that Myanmar was drafting its National AI Strategy and Policy[151] as of February 2025.

Agentic AI security: Not directly addressed

AI-enabled social engineering: Not directly addressed

New Zealand

New Zealand launched its National AI Strategy[152] and Responsible guidance for businesses[153] in July 2025. Its focus is on increasing adoption and innovation in practical applications, with a clear strategic and economic rationale for AI adoption. It sets commitments to stable, enabling policy; government support to demystify AI for business; and building an AI-ready workforce.

Agentic AI security: Not directly addressed

AI-enabled social engineering: Not directly addressed

Philippines

The Philippines launched the National AI Strategy and Roadmap (NAISR) 2.0[154] in 2024, in tandem with the launch of the Center for AI Research (CAIR). Helmed by the Department of Trade and Industry (DTI), the NAISR has a "whole-of-government" approach with six objectives: (1) increase the regional and global competitiveness of local industries through AI-driven industrial growth; (2) identify key areas for investment in R&D and technology application to advance new processes, products, and services; (3) promote triple-helix collaborations in R&D, crucial for national development; (4) prepare the workforce for future jobs; (5) attract major industries to create jobs in the Philippines; and (6) ensure the responsible rollout and governance of AI technologies—emphasizing ethics, data privacy, and minimizing the negative impacts on society—by creating regulatory frameworks to guide AI deployment, promoting transparency and accountability, and fostering public awareness.

There are a number of AI related bills[155] filed in the current Congress, with the passage of an AI regulatory framework being one of the priorities of the present government. A regional AI-governance framework under the Association of Southeast Asian Nations (ASEAN) is also under consideration, with the Philippines set to chair ASEAN in 2026.

Agentic AI security: Not directly addressed

AI-enabled social engineering: Not directly addressed

Singapore

Singapore has emerged as a proactive leader in AI security governance, with the Cyber Security Agency (CSA) releasing comprehensive AI security guidelines, although they were only partially directed at emerging agentic AI security concerns.

In October 2024, the CSA published its Guidelines on Securing AI Systems,[156] alongside a community-driven Companion Guide on Securing AI Systems,[157] which was updated in May 2025 with clarifications and expanded examples on adversarial robustness testing and secure model retraining. These guidelines emphasize a lifecycle approach to AI security, advocating for "secure by design" and "secure by default" principles while addressing both classical cybersecurity risks and novel adversarial machine learning threats. While the frameworks do not specifically focus on agentic AI systems, they establish foundational security principles that can be extended—at least in part—to autonomous agent deployments.

More recently, Singapore's Government Technology Agency published an Agentic AI Primer [158]in April 2025 to provide clear implementation frameworks for AI agents, noting briefly that autonomous AI systems are a distinct category requiring specialized governance approaches, although with no specific references to security aspects.

Agentic AI security: Not directly addressed

AI-enabled social engineering: Not directly addressed

South Korea

Korea is a regional leader in AI policy development and established its National Strategy for Artificial Intelligence[159] in 2019. The strategy's vision is "Toward AI World Leader beyond IT – AI for Everyone, AI of Everything". Its core strategies and goals are: (1) innovation of AI competitiveness through AI infrastructure enhancement, securing competitiveness in AI technology, bold regulatory innovation and law revision, and nurturing global AI start-ups; (2) full scale AI utilization by nurturing world-class AI talent and educating the public, diffusing AI technology across all industries, and building

a best-in-class digital government; and (3) harmony and coexistence with AI through establishing an inclusive job safety network, preventing dysfunction, and establishing AI ethics.

South Korea's AI Framework Act,[160] effective January 2026, integrates both ethics-oriented soft laws and hard regulatory obligations to ensure AI safety and security. The Act mandates AI operators, especially those managing "high-impact AI," to implement comprehensive risk management systems, conduct impact assessments on fundamental human rights, and maintain transparency through clear user notifications and explanations of AI-generated results. The Ministry of Science and ICT (MSIT) oversees compliance and has investigative and enforcement powers, including fines. Foreign AI providers must appoint domestic representatives for accountability. The regulation aims to balance innovation with protective measures against AI risks, emphasizing safety, reliability, and human oversight.

Agentic AI security: Not directly addressed

AI-enabled social engineering: Not directly addressed

Thailand

The National Cyber Security Agency launched Thailand's first AI Security Guidelines: Secure and Responsible Use of AI[161] on October 1, 2025. The Guidelines are non-binding and provide a comprehensive set of principles for developing secure AI systems, including (1) identifying threats and vulnerabilities, (2) implementing robust security frameworks, and (3) governance and risk management approaches.

Among those under consideration, Thailand is the first country to have addressed emerging AI security threats. The Guidelines identify several emerging threats posed by agentic AI, including multi-step attacks involving system scans, backdoor deployments, and ransomware releases. They also highlight the risks of unauthorized or harmful actions caused by malfunctioning AI agents. Recommended control measures include kill chain monitoring, Security Orchestration, Automation and Response (SOAR) systems, autonomous defense agents, digital deception tools, regulated Software Bills of Materials (SBOM), CI/CD pipeline protection, and adherence to AI and information security management systems standards such as ISO/IEC 42001 and ISO/IEC 2700.

The Guidelines also include an AI security checklist that outlines specific security requirements for Agentic AI across the design, verification, and operational phases. In the design phase, developers must define and limit the action space of Agentic AI based on the least privilege principle. In the verification phase, penetration tests should be conducted to determine whether AI agents can be manipulated to perform unauthorized or malicious actions. While operational, continuous monitoring and comprehensive logging of all automated actions are required to enable retrospective auditing.

On AI-enabled phishing, the Guidelines briefly recognize it as a sophisticated cybersecurity risk. They address this through explicit prohibitions and testing procedures, recommending a misuse testing phase to assess whether the AI system could generate harmful outputs, such as phishing emails, to verify the effectiveness of safety mechanisms and guardrails.

Agentic AI security: Not directly addressed

AI-enabled social engineering: Not directly addressed

Timor Leste

Timor Leste does not have a specific AI law as of September 2025. However, in line with the country's upcoming accession into ASEAN (expected in Oct 2025), the country has been participating in regional discussions, such as the UNESCO AI Readiness Assessment[162] to build ethical frameworks for AI 2025, and the ASEAN AI Summit 2025 (AAIMS 25).[163] Timor Leste is expected to develop its technology capabilities rapidly as the country continues to develop.

Agentic AI security: Not directly addressed

AI-enabled social engineering: Not directly addressed

Vietnam

In September 2025, Vietnam's Ministry of Science and Technology announced that the country will issue an updated version of its National strategy for AI research and application[164] (Decision No. 127/QD-TTg of 2021 by the Prime Minister), first approved[165] in 2021.[166] The strategy is to position AI in a way that leverages the technology for economic growth, social development, and global competitiveness. It proposes the principle of "AI for humans – safe, autonomous, cooperative, inclusive, and sustainable." Six core principles form this legislation: (1) risk-based regulation, (2) transparency and accountability, (3) human-centric development, (4) domestic AI autonomy, (5) AI as a driver of sustainable growth, and (6) digital sovereignty, with data, infrastructure, and AI technology being three strategic pillars.

Agentic AI security: Not directly addressed

AI-enabled social engineering: Not directly addressed

Endnotes

1 https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai

2 https://www-cdn.anthropic.com/b2a76c6f6992465c09a6f2fce282f6c0cea8c200.pdf

3 https://blog.knowbe4.com/key-takeaways-from-the-2025-phishing-threat-trends-report

4 https://www.proofpoint.com/us/blog/threat-insight/cybercriminals-abuse-ai-website-creation-app-phishing

5 https://www.ibm.com/think/x-force/ai-vs-human-deceit-unravelling-new-age-phishing-tactics

6 https://www.ibm.com/think/topics/ai-security

7 https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai

8 https://ff.co/ai-statistics-trends-global-market/

9 https://kpmg.com/mt/en/home/media/press-releases/2025/05/global-study-reveals-tension-between-ai-benefits-and-risks-and-highlights-a-governance-gap.html

10 https://ff.co/ai-statistics-trends-global-market/

11 https://online.hbs.edu/blog/post/benefits-of-ai-in-business

12 https://ccapac.asia/wp-content/uploads/2024/10/CCAPAC-2024-Report.pdf

13 https://www.cisco.com/c/en/us/products/security/state-of-ai-security.html

14 https://home.treasury.gov/system/files/136/G7-Cyber-Expert-Group-Statement-AI-and-Cybersecurity-2025.pdf

15 https://www.cxtoday.com/crm/a-customer-service-ai-agent-spits-out-complete-salesforce-records-in-an-attack-by-security-researchers/

16 https://openai.com/global-affairs/disrupting-malicious-uses-of-ai-october-2025/

17 https://www.cisco.com/c/en/us/products/security/state-of-ai-security.html#:~:text=The%20State%20of%20AI%20Security%20report%20will%20cover%3A,partnership%20agreements%2C%20and%20security%20frameworks.

18 https://engage.checkpoint.com/2025-ai-security-report

19 https://www.cisco.com/c/en/us/products/security/state-of-ai-security.html#:~:text=The%20State%20of%20AI%20Security%20report%20will%20cover%3A,partnership%20agreements%2C%20and%20security%20frameworks.

20 https://www.crowdstrike.com/explore/2025-global-threat-report-en-gb?utm_medium=org

21 https://unit42.paloaltonetworks.com/agentic-ai-threats/

22 https://www.enisa.europa.eu/publications/enisa-threat-landscape-2025

23 https://www.ncsc.gov.uk/report/impact-ai-cyber-threat-now-2027

24 https://www.csa.gov.sg/resources/publications/singapore-cyber-landscape-2024-2025

25 https://www.helpnetsecurity.com/2025/05/30/ai-agents-organizations-risk/

26 https://arxiv.org/html/2505.10468v1

27 https://www.cisco.com/c/en/us/products/collateral/networking/software/crosswork-network-automation/crosswork-multi-agentic-ai-framework-aag.html

28 https://www.salesforce.com/agentforce/how-it-works/

29 https://arxiv.org/html/2508.20866v1

30 https://www.sciencedirect.com/science/article/pii/S0308596125000734

31 https://www.securityjourney.com/post/experts-reveal-how-agentic-ai-is-shaping-cybersecurity-in-2025

32 https://www.sciencedirect.com/science/article/pii/S0308596125000734

33 https://www.scworld.com/feature/ai-to-change-enterprise-security-and-business-operations-in-2025

34 https://www.coalitionforsecureai.org/announcing-the-cosai-principles-for-secure-by-design-agentic-systems/#:~:text=2025%20has%20seen%20the%20launch,agent

35 https://genai.owasp.org/resource/state-of-agentic-ai-security-and-governance-1-0/

36 https://genai.owasp.org/resource/state-of-agentic-ai-security-and-governance-1-0/

37 https://www.sailpoint.com/press-releases/sailpoint-ai-agent-adoption-report

38 https://www-cdn.anthropic.com/b2a76c6f6992465c09a6f2fce282f6c0cea8c200.pdf

39 https://unit42.paloaltonetworks.com/2025-unit-42-global-incident-response-report-social-engineering-edition/

40 https://blog.knowbe4.com/70-to-90-of-all-malicious-breaches-are-due-to-social-engineering-and-phishing-attacks

41 https://blog.knowbe4.com/key-takeaways-from-the-2025-phishing-threat-trends-report

42 https://www.aon.com/cyber-risk-report/asia-pacifics-commitment-to-cyber-security-pays-off

43 https://csrc.nist.gov/glossary/term/social_engineering

44 Traditional social engineering techniques include pretexting (creating fabricated scenarios), diversion theft (redirecting deliveries or communications), water holing (compromising frequently visited websites), baiting (offering incentives for malicious actions), quid-pro-quo arrangements (offering services in exchange for information), tailgating (physical following), and honey trap operations (romantic or professional relationship exploitation). Source: https://www.knowbe4.com/what-is-social-engineering

45 https://csrc.nist.gov/glossary/term/phishing

46 https://blog.knowbe4.com/ai-attacks-are-coming-in-a-big-way-now

47 https://www.itpro.com/security/cyber-attacks/watch-out-for-fake-zoom-invites-hackers-are-abusing-connectwise-screenconnect-to-take-over-devices

48 https://www.ibm.com/think/x-force/ai-vs-human-deceit-unravelling-new-age-phishing-tactics

49 https://blog.knowbe4.com/ai-attacks-are-coming-in-a-big-way-now

50 https://blog.knowbe4.com/ai-attacks-are-coming-in-a-big-way-now

51 https://blog.knowbe4.com/ai-attacks-are-coming-in-a-big-way-now

52 https://www.ibm.com/think/topics/ai-security

53 The following list of industry solutions is non-exhaustive. For additional information on products addressing LLM and Generative AI risks, please refer to OWASP's AI Security Solutions Landscape. https://genai.owasp.org/ai-security-solutions-landscape/

54 https://www.securityjourney.com/post/experts-reveal-how-agentic-ai-is-shaping-cybersecurity-in-2025

55 https://systemweakness.com/beyond-sandboxes-layered-security-for-ai-agent-infrastructure-e9d25c8235c8

56 https://www.aisi.gov.uk/work/the-inspect-sandboxing-toolkit-scalable-and-secure-ai-agent-evaluations

57 https://www.getmonetizely.com/articles/how-can-we-secure-agentic-ai-systems-against-emerging-threats

58 https://arxiv.org/html/2505.02077v1

59 https://kenhuangus.substack.com/p/zero-trust-security-for-multi-agent

60 https://newsroom.cisco.com/c/r/newsroom/en/us/a/y2025/m06/cisco-transforms-security-for-the-agentic-ai-era-further-fusing-security-into-the-network.html#:~:text=Enabling%20Agentic%20AI%20Securely%3A%20The,Universal%20Zero%20Trust%20architecture%20to

61 https://newsroom.cisco.com/c/r/newsroom/en/us/a/y2025/m06/cisco-transforms-security-for-the-agentic-ai-era-further-fusing-security-into-the-network.html#:~:text=Enabling%20Agentic%20AI%20Securely%3A%20The,Universal%20Zero%20Trust%20architecture%20to

62 https://www.lasso.security/blog/agentic-ai-tools

63 https://arxiv.org/html/2505.02077v1

64 https://www.arionresearch.com/blog/g9jiv24e3058xsivw6dig7h6py7wml

65 https://www.legitsecurity.com/aspm-knowledge-base/ai-security-posture-management

66 https://www.dsalta.com/resources/articles/why-ai-agents-need-compliance-too-managing-risk-in-the-age-of-autonomous-systems

67 https://cybermagazine.com/articles/knowbe4-launches-ai-agents-to-counter-phishing-threats

68 https://securitybrief.co.uk/story/knowbe4-s-ai-platform-helps-cut-phishing-risk-to-just-one-percent

69 https://www.knowbe4.com/products/aida

70 https://securitybrief.co.uk/story/knowbe4-s-ai-platform-helps-cut-phishing-risk-to-just-one-percent

71 https://www.ciscoinvestments.com/protecting-digital-identity-deepfake-invests-getreal-security

72 https://my.idc.com/getdoc.jsp?containerId=US53703125

73 https://controld.com/blog/phishing-statistics-industry-trends/

74 https://www.knowbe4.com/products/cloud-email-security

75 https://arxiv.org/abs/2505.15233

76 https://pubmed.ncbi.nlm.nih.gov/37367470/

77 https://www.sciencedirect.com/science/article/pii/S1077314224002248

78 https://www.digital.gov.au/sites/default/files/documents/2024-08/Policy%20for%20the%20responsible%20use%20of%20AI%20in%20government%20v1.1.pdf

79 https://www.industry.gov.au/publications/voluntary-ai-safety-standard

80 The Five Eyes countries are Australia, Canada, New Zealand, the United Kingdom, and the United States.

81 https://www.cyber.gov.au/business-government/secure-design/artificial-intelligence/deploying-ai-systems-securely

82 https://media.defense.gov/2025/May/22/2003720601/-1/-1/0/CSI_AI_DATA_SECURITY.PDF

83 https://www.protectivesecurity.gov.au/

84 https://www.dta.gov.au/articles/new-ai-technical-standard-support-responsible-government-adoption

85 https://www.itnews.com.au/news/sa-gov-establishes-office-for-ai-619064

86 https://www.security.sa.gov.au/__data/assets/pdf_file/0009/1167723/SA-Cyber-Security-Strategy-2025-2031.pdf

87 https://drive.ncsa.or.th/s/fpp25icTY86JxLn?dir=/&editing=false&openfile=true

88 https://isomer-user-content.by.gov.sg/36/e05d8194-91c4-4314-87d4-0c0e013598fc/Guidelines%20on%20Securing%20AI%20Systems.pdf

89 https://isomer-user-content.by.gov.sg/36/3cfb3cd5-0228-4d27-a596-3860ef751708/Companion%20Guide%20on%20Securing%20AI%20Systems.pdf

90 https://www.developer.tech.gov.sg/guidelines/standards-and-best-practices/agentic-ai-primer.html

91 This list is non exhaustive. For a full overview of current standardization efforts, please refer to the U.K.'s (https://www.gov.uk/government/publications/ai-cyber-security-code-of-practice) Implementation guide for the AI Cyber Security Code of Practice (January 2025) (https://genai.owasp.org/resource/state-of-agentic-ai-security-and-governance-1-0/) and the more recent OWASP's publication State of Agentic AI Security and Governance (August 2025).

92 https://www.iso.org/standard/42006

93 https://www.iso.org/standard/42005

94 https://www.iso.org/standard/42001

95 https://www.etsi.org/deliver/etsi_ts/104200_104299/104223/01.01.01_60/ts_104223v010101p.pdf

96 https://www.etsi.org/deliver/etsi_tr/104100_104199/104128/01.01.01_60/tr_104128v010101p.pdf

97 https://csrc.nist.gov/pubs/ai/100/2/e2025/final

98 https://www.nist.gov/news-events/news/2025/08/nist-releases-control-overlays-securing-ai-systems-concept-paper#:~:text=The%20concept%20paper%20outlines%20proposed,and%20controls%20for%20AI%20developers.

99 https://csrc.nist.gov/pubs/sp/800/53/r5/upd1/final

100 https://www.nist.gov/itl/ai-risk-management-framework

101 https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai

102 https://ec.europa.eu/newsroom/dae/redirection/document/118119

103 https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/Publications/ANSSI-BSI-joint-releases/LLM-based_Systems_Zero_Trust.pdf?__blob=publicationFile&v=3

104 https://www.gov.uk/government/publications/ai-cyber-security-code-of-practice

105 https://assets.publishing.service.gov.uk/media/679cae441d14e76535afb630/Implementation_Guide_for_the_AI_Cyber_Security_Code_of_Practice.pdf

106 https://www.aisi.gov.uk/

107 https://www.aisi.gov.uk/blog/examining-backdoor-data-poisoning-at-scale

108 https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf

109 https://www.congress.gov/bill/119th-congress/senate-bill/146

110 https://www.mitre.org/news-insights/news-release/mitre-launches-ai-incident-sharing-initiative

111 https://incidentdatabase.ai/

112 https://genai.owasp.org/

113 https://www.coalitionforsecureai.org/

114 https://atlas.mitre.org/

115 https://cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/msc/documents/presentations/CSR/International-cybersecurity-regulatory-alignment.pdf

116 https://threatbeat.com/welcome-to-the-sandbox-championing-cyber-resilience-over-regulatory-theater/

117 https://www.csa.gov.sg/news-events/press-releases/singapore-signs-mutual-recognition-arrangements-with-republic-of-korea-and-germany-on-cybersecurity-labelling-for-consumer-smart-products

118 https://reports.weforum.org/docs/WEF_Global_Cybersecurity_Outlook_2025.pdf

119 https://www.digital.gov.au/sites/default/files/documents/2024-08/Policy%20for%20the%20responsible%20use%20of%20AI%20in%20government%20v1.1.pdf

120 https://www.industry.gov.au/publications/voluntary-ai-safety-standard

121 The Five Eyes countries are Australia, Canada, New Zealand, the United Kingdom, and the United States.

122 https://www.cyber.gov.au/business-government/secure-design/artificial-intelligence/deploying-ai-systems-securely

123 https://media.defense.gov/2025/May/22/2003720601/-1/-1/0/CSI_AI_DATA_SECURITY.PDF

124 https://www.protectivesecurity.gov.au/

125 https://www.dta.gov.au/articles/new-ai-technical-standard-support-responsible-government-adoption

126 https://www.itnews.com.au/news/sa-gov-establishes-office-for-ai-619064

127 https://www.security.sa.gov.au/__data/assets/pdf_file/0009/1167723/SA-Cyber-Security-Strategy-2025-2031.pdf

128 https://www.aiti.gov.bn/regulatory/ai-governance-and-ethics/

129 https://thescoop.co/2025/06/24/ai-will-be-central-to-next-digital-master-plan/

130 https://kiripost.com/stories/cambodia-moves-cautiously-towards-ai-regulation-amid-rising-concerns

131 https://www.unesco.org/en/articles/cambodia-launches-ai-readiness-assessment-report-guide-ethical-and-inclusive-digital-transformation

132 https://digichina.stanford.edu/work/full-translation-chinas-new-generation-artificial-intelligence-development-plan-2017/

133 https://www.gov.cn/xinwen/2017-11/15/content_5239965.htm

134 https://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm

135 https://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm

136 https://ai-innovation.id/

137 https://money.kompas.com/read/2025/09/25/121738626/indonesia-bakal-terbitkan-dua-perpres-ai-sekaligus-apa-isinya

138 https://www.digitalindia.gov.in/

139 https://indiaai.gov.in/

140 https://indiaai.gov.in/hub/indiaai-futureskills

141 https://www.skillindiadigital.gov.in/

142 https://www.pib.gov.in/PressReleasePage.aspx?PRID=2154268

143 https://www8.cao.go.jp/cstp/ai/aistratagy2022en.pdf

144 https://www.soumu.go.jp/hiroshimaaiprocess/en/index.html

145 https://www.kojimalaw.jp/wp/wp-content/uploads/2025/09/Japan-AI-Promotion-Act-KOJIMA-LAW-OFFICES-jp-en-reference-translation.pdf

146 https://asianews.network/ai-seen-to-advance-digital-governance-industry-ethics-in-laos/

147 https://laopdr.un.org/en/300785-remarks-capacity-building-programme-ai-governance

148 https://hkifoa.com/wp-content/uploads/2024/12/ai-roadmap-2025-malaysia.pdf

149 https://www.thestar.com.my/news/nation/2025/08/05/govt-developing-ai-guidelines-to-combat-online-scam-to-be-ready-early-2026

150 https://www.myanmaritv.com/news/ai-strategy-and-policy-meeting-drafting-nas-and-nap-draft

151 https://english.news.cn/20250219/7be9255fc3b14a1796ae6919a693b171/c.html

152 https://www.mbie.govt.nz/assets/new-zealands-strategy-for-artificial-intelligence.pdf

153 https://www.mbie.govt.nz/business-and-employment/business/support-for-business/responsible-ai-guidance-for-businesses

154 https://erikalegara.com/uploads/NAISR2.0_July2024.pdf

155 https://www.reuters.com/technology/philippines-propose-asean-ai-regulatory-framework-house-speaker-says-2024-01-17/

156 https://isomer-user-content.by.gov.sg/36/e05d8194-91c4-4314-87d4-0c0e013598fc/Guidelines%20on%20Securing%20AI%20Systems.pdf

157 https://isomer-user-content.by.gov.sg/36/3cfb3cd5-0228-4d27-a596-3860ef751708/Companion%20Guide%20on%20Securing%20AI%20Systems.pdf

158 https://www.developer.tech.gov.sg/guidelines/standards-and-best-practices/agentic-ai-primer.html

159 https://www.msit.go.kr/bbs/view.do?sCode=eng&nttSeqNo=9&bbsSeqNo=46&mId=10

160 https://www.law.go.kr/%EB%B2%95%EB%A0%B9/%EC%9D%B8%EA%B3%B5%EC%A7%80%EB%8A%A5%20%EB%B0%9C%EC%A0%84%EA%B3%BC%20%EC%8B%A0%EB%A2%B0%20%EA%B8%B0%EB%B0%98%20%EC%A1%B0%EC%84%B1%20%EB%93%B1%EC%97%90%20%EA%B4%80%ED%95%9C%20%EA%B8%B0%EB%B3%B8%EB%B2%95/(20676,20250121)

161 https://drive.ncsa.or.th/s/fpp25icTY86JxLn?dir=/&editing=false&openfile=true

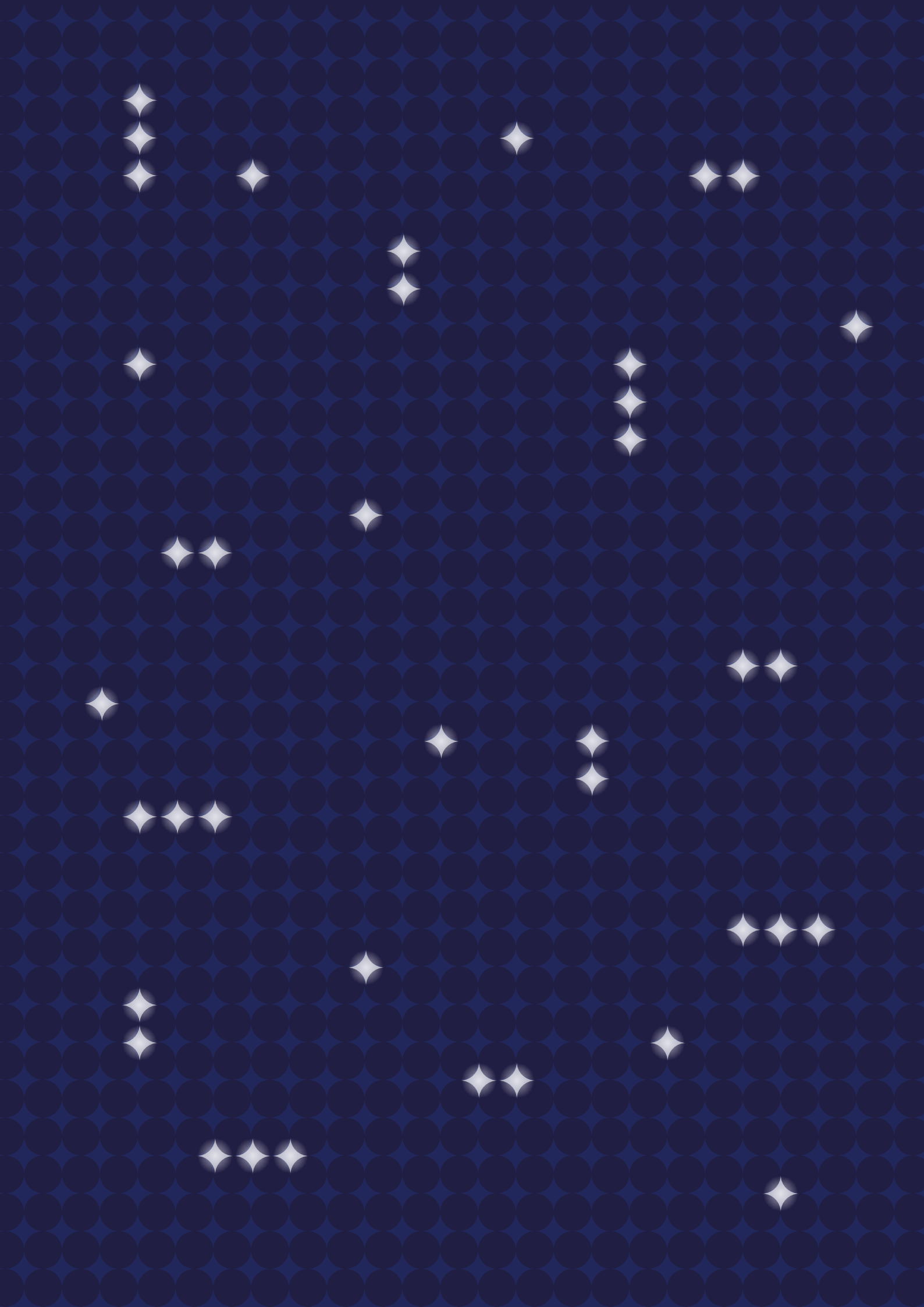
162 https://www.unesco.org/en/articles/preparing-timor-leste-embrace-artificial-intelligence

163 https://timor-leste.gov.tl/?p=44825&lang=en&n=1

164 https://www.tilleke.com/insights/vietnams-ai-push-updated-national-strategy-and-first-ai-law-by-end-of-2025/8/

165 https://research.csiro.au/aus4innovation/wp-content/uploads/sites/578/2025/04/3.-National-strategy-for-Artificial-Intelligence-2030-ENG.pdf

166 https://vietnamlawmagazine.vn/national-strategy-for-ai-research-and-application-approved-27594.html



The Coalition for Cybersecurity in Asia-Pacific or CCAPAC is a group of dedicated industry stakeholders who are working to positively shape the cybersecurity environment in Asia through policy analysis, engagement, and capacity building.

ccapac.asia