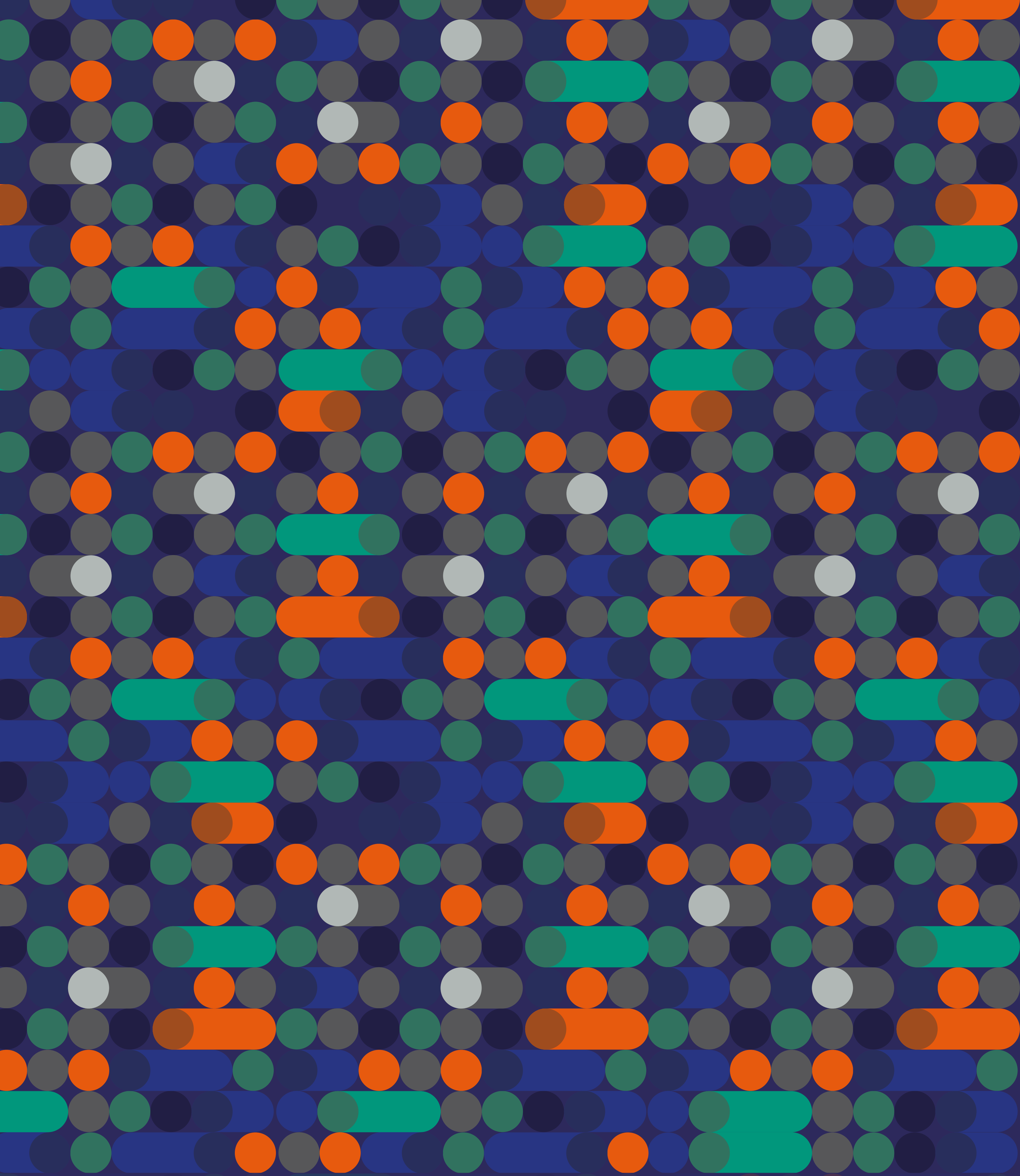
2024

CCAPAC Report:

AI and Cybersecurity

https://ccapac.asia

COALITION FOR
CYBERSECURITY
IN ASIA-PACIFIC



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

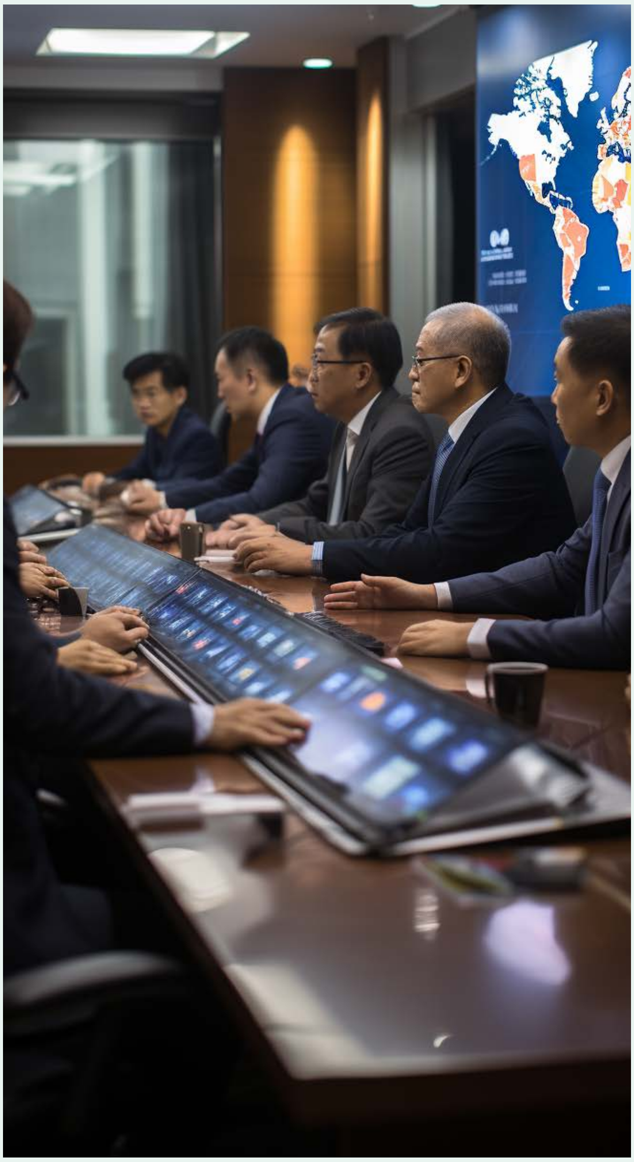
Table of Contents

Executive Summary

Artificial Intelligence (AI) is becoming a pivotal force in driving innovation across various sectors, including healthcare, finance, transportation, and manufacturing. However, as AI is increasingly integrated into critical systems and infrastructures, it introduces significant cybersecurity risks that must be effectively managed. This report provides a comprehensive analysis of these risks and offers strategic guidance for developing robust policies to mitigate them in the Asia-Pacific region.

The report begins by exploring the current AI landscape, categorizing AI systems into predictive and generative types, and detailing their applications across key sectors. While AI holds the potential to deliver transformative advancements, it also brings unique vulnerabilities, such as data poisoning, model evasion attacks, and ethical concerns, which could compromise the integrity and security of AI systems. To address these challenges, we propose a holistic framework for AI cybersecurity. This framework includes key components such as oversight, lifecycle management, model security, data governance, transparency, and incident response strategies. It is designed to integrate seamlessly with existing laws and internationally recognized standards, ensuring a cohesive and effective approach to AI governance across the Asia-Pacific region.

Our policy recommendations emphasize the need to update national cybersecurity strategies, establish AI-specific guidelines, invest in research and development, foster international cooperation, and promote AI literacy. These measures are crucial to ensure that the deployment of AI systems is both secure and ethical, supporting innovation while proactively addressing emerging threats.



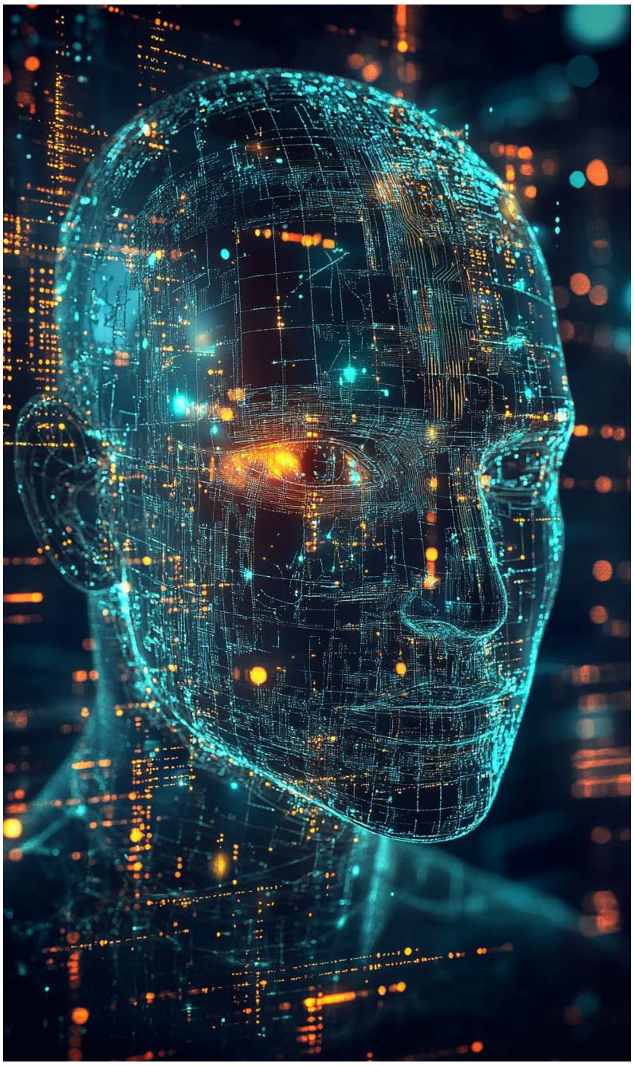
The importance of a coordinated effort among governments, industries, and academia to develop and implement robust AI cybersecurity policies cannot be understated. Such policies are essential not only for protecting against current risks but also for anticipating and mitigating future challenges, ensuring that AI continues to be a driving force for positive change in the region.

Chapter 1
Introduction

AI has emerged as one of the most transformative technologies of our time, with the capability to revolutionize industries, reshape societies, and drive unprecedented innovation across the globe. From healthcare and finance to transportation and energy, AI is being leveraged to solve complex problems, optimize processes, and create new opportunities for growth and development. Rapid advancements in AI technologies, such as machine learning, deep learning, and natural language processing, and most recently, generative AI tools made available for consumer use, have enabled the development of intelligent systems that can perform tasks once exclusively within the domain of human intelligence.

However, as AI becomes increasingly integrated into our daily lives and infrastructure, it is essential to recognize and address the cybersecurity risks associated with this technology. The very features that make AI so transformative–its ability to learn, adapt, and make decisions autonomously–also introduce new vulnerabilities, attack surfaces, and attack capabilities that malicious actors can exploit. These risks are compounded by the increasing complexity and opacity of AI systems, which can make it difficult to detect and mitigate potential security breaches.

The rapid adoption of AI has highlighted the need for comprehensive and aligned policies to address potential cybersecurity risks and ensure the safe, secure, and ethical deployment of AI systems.[1] Governments, industry, and academia must work together to establish a robust governance framework that addresses the unique challenges posed by AI while fostering innovation and promoting the responsible development and deployment of this transformative technology.



This report aims to provide a comprehensive understanding of AI cybersecurity risks as an informed tool to aid effective policymaking. It offers an overview of the current AI landscape, including key technologies, applications, and standards, and provides policy recommendations for governments and regulators to strengthen and align policies for the secure deployment of AI systems in the Asia-Pacific region.

Chapter 2
The Current AI Landscape

2.1 Overview of AI technologies and their applications

While the term "artificial intelligence" has become commonplace, there is still no universally accepted definition. For instance, the United Kingdom's National Cyber Security Centre (NCSC) defines AI as "any computer system that can perform tasks usually requiring human intelligence, such as visual perception, text generation, speech recognition, or translation between languages."[2] This definition underscores the broad and diverse nature of AI technologies. In contrast, the definition provided by the Association for the Advancement of Artificial Intelligence (AAAI) describes AI as "the scientific understanding of the mechanisms underlying thought and intelligent behavior and their embodiment in machines."[3] This definition emphasizes the scientific and cognitive aspects of AI, focusing on understanding and replicating human intelligence in machines.

For the purposes of this policy-orientated report, we will use the definition provided by the European Union (EU) AI Act, Article 3.1, as a foundational reference. According to this definition, an artificial intelligence system is "software developed with one or more techniques and approaches that can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with".[4] We adopt this definition as it encapsulates the multifaceted nature of AI systems, and for cybersecurity policymaking, it provides a foundation for understanding the diverse branches of AI and their relevant applications. This understanding is useful for navigating AI standards and policies, as we will elaborate below.

2.1.1 Stages of AI Development

Though the field of AI has seen rapid evolution in recent years, the general concept was developed in 1956 during the Summer Research Project on Artificial Intelligence workshop at Dartmouth.[5] Over the years, there have been promising advances and setbacks in the development and use of AI. Some AI techniques, such as machine learning, became practical in the 1990s, taking off in the late 2000s and early 2010s with the explosion of data and computing power.[6] Generative AI gained traction in 2014 and gained significant public attention in 2020 with the introduction of ChatGPT 3.0 by OpenAI.[7] Despite these advancements, many experts believe we are still in the early stages of AI development. AI capability is categorized into two stages of development:

- **Artificial Narrow Intelligence (ANI):** ANI systems perform specific tasks or a narrow range of tasks. Examples include virtual personal assistants and recommendation systems. Most current AI capabilities fall into this category.[8]
- **Artificial General Intelligence (AGI):** Also known as Strong AI, AGI has the potential for software to have human-like intelligence, self-learning abilities, and a broad set of adaptable functions. As such, AGI software is expected to be able to perform novel tasks which they have not been trained for. However, AGI is currently a hypothetical technology at the forefront of research.

2.1.2 Two Classes of AI Systems: Predictive and Generative

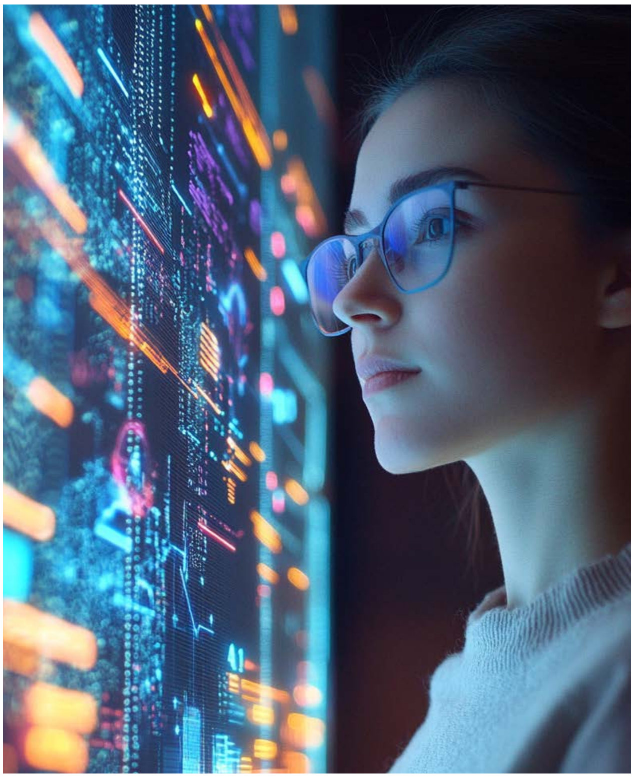
AI, particularly in the current narrow sense, is broadly categorized into two classes: Predictive AI and Generative. Understanding these distinctions is crucial for policymakers to effectively address the unique challenges and opportunities presented by each class.

Predictive AI systems are designed to analyze data and make forecasts about future events. These systems are widely used in various fields such as in finance for market predictions, healthcare for disease outbreak forecasting, logistics for supply chain optimization, and environment for climate prediction. For example, in the healthcare sector, IBM's Watson Health employs predictive analytics to improve patient care by forecasting disease progression and recommending personalized treatment plans.[9] Cisco's Predictive Network technology gathers data from a range of telemetry sources, and leverages AI and models to learn patterns, predict user experience issues, and provide problem solving options, thus creating self-healing networks that can learn, predict and plan.[10] In the environmental sector, particularly within ASEAN, where over 50% of global disaster mortalities occurred between 2004 to 2014,[11] predictive AI can play a pivotal role in disaster management. It can forecast the paths of typhoon and assess the likelihoods of earthquakes and tsunamis, enabling more effective response strategies.

Generative AI systems, on the other hand, are capable of creating new content, such as text, images, audio and video, based on the data they have been trained on. The recent surge in interest and development in Generative AI has been driven by advancements in techniques such as Generative Adversarial Networks (GANs) and Large Language Models (LLM). OpenAI's ChatGPT platform uses this class of AI to generate coherent text, images, and code based on user prompts. Generative AI can also be leveraged for creating digital art, literature, and music reflective of cultural heritage, as well as building intelligent interactive assistants.

2.1.3 Types of AI

AI is a broad term encompassing diverse technologies, each utilizing distinct algorithms and training methods tailored to specific applications.



Machine Learning (ML), a subset of AI, focuses on developing algorithms that enable computers to learn from data and improve their performance over time without explicit programming. ML is predominantly used for predictive tasks, such as weather forecasting, stock market analysis, spam detection, and image recognition.

Deep Learning (DL), a specialized branch of ML, utilizes multi-layered neural networks to analyze complex data structures. DL excels at tasks like image and speech recognition due to its ability to automatically extract features from raw data. While it can be applied to traditional ML tasks, DL also enables more advanced capabilities such as LLMs, Computer Vision, Autonomous Systems, and Medical Diagnosis

2.2 AI application across key sectors

AI has evolved significantly since its inception in the mid-20th century. Since the infamous victory of the AI system Deep Blue over the world chess champion Gary Kasparov in 1997,[12] the technology has witnessed various phases of development, each characterized by different approaches and breakthroughs – leading to wider application. Today, use cases span across multiple key sectors and stand to advance national goals in various forms.

2.2.1 Government and the Public Sector

AI can transform the government and public sector by improving service delivery, enhancing decision-making, and optimizing resource allocation. Deep learning and natural language processing (NLP) are applied to large datasets, such as electronic health records and public feedback, to support clinical decision-making, detect suspicious activities, and automate administrative tasks. For example, in Singapore, the government uses AI in its Smart Nation initiative, which aims to enhance urban living through technology. The MyResponder app, powered by AI, helps to locate the nearest trained responders in cases of cardiac emergencies, significantly improving response times and survival rates.[13]

Additionally, AI can monitor environmental conditions, predict weather events, and manage emergency situations. An increasingly useful function is the ability of AI models to analyze satellite imagery and meteorological data to predict natural disasters like floods or earthquakes, allowing for timely evacuation and resource deployment.[14] Governments are also leveraging AI to boost the efficiency of administrative operations, ensure compliance with regulatory standards, and improve economic analysis, trade surveillance, and infrastructure planning through predictive maintenance and procurement analytics. For example, predictive analytics can forecast infrastructure wear and tear, enabling timely maintenance, and reducing costs.

2.2.2 Healthcare

AI can revolutionize healthcare by improving diagnosis, treatment, and patient care. AI techniques, such as DL and NLP, are applied to electronic health records, medical literature, and genomic data to identify patterns, predict outcomes, and support clinical decision-making. As AI algorithms analyze medical images to detect early signs of diseases such as cancer, many lives can be saved. Beyond saving lives, AI can improve the quality of life by aiding in the development of personalized treatment plans, through analyzing a patient's genetic information and medical history to recommend the most effective treatments.

In drug discovery, AI accelerates the process by predicting how different compounds will interact with targets in the human body, potentially reducing the time and cost involved in bringing new drugs to market. Moreover, innovators in the field have their sights set on perfecting robotic surgery and AI-assisted medical devices to enhance surgical precision and outcomes, reducing the risk of complications, and speeding up recovery times.

2.2.3 Finance

AI offers significant benefits in the financial industry, from fraud detection and risk assessment to algorithmic trading and robo-advisory services. Machine learning models analyze vast amounts of financial data to identify fraudulent transactions, assess credit risk, and detect money laundering. For example, AI can flag unusual patterns in transactions that might indicate fraud, allowing for quicker responses and reducing financial losses.

In the Philippines, AI-powered robo-advisors provide personalized investment advice, making financial planning more accessible to a broader population. According to the Bangko Sentral ng Pilipinas, these AI-driven services have increased financial inclusion by offering low-cost advisory services to underserved communities.[15]

2.2.4 Transportation

AI is central to the development of autonomous vehicles, which promise to revolutionize transportation by improving safety, reducing congestion, and enhancing mobility. Techniques such as computer vision and deep reinforcement learning enable vehicles to perceive their environment, make decisions, and navigate complex traffic scenarios. Whilst human supervision is still necessary, autonomous vehicles can reduce accidents caused by human error, improve fuel efficiency, and provide mobility solutions for people unable to drive, such as those living with disabilities. In more developed economies, AI is used to optimize traffic management by analyzing real-time data from traffic cameras, sensors, and GPS devices to adjust traffic light timings and manage congestion. Predictive analytics can forecast demand for public transportation services, allowing for better scheduling and resource allocation. In logistics, AI improves route planning, reduces delivery times, and lowers operational costs, enhancing overall efficiency and customer satisfaction.

2.2.5 Energy

AI optimizes energy production, distribution, and consumption. For example, in Thailand, machine learning models predict energy demand and optimize renewable energy integration. The Electricity Generating Authority of Thailand (EGAT) uses AI to analyze weather patterns and forecast solar and wind energy production, helping to balance supply and demand on the grid.[16] In the oil and gas industry, AI optimizes drilling operations by analyzing geological data to identify the most promising drilling sites, predict equipment failures, and improve safety. Predictive maintenance models can forecast when equipment is likely to fail, allowing for timely interventions that prevent costly downtime and accidents.

2.2.6 Manufacturing

AI enables predictive maintenance, quality control, supply chain optimization, and robotics integration.

Sensor data from equipment can be analyzed to predict and prevent failures, reducing downtime and maintenance costs. For instance, AI can predict when a machine part is likely to wear out and schedule maintenance before it breaks down, avoiding production delays. Computer vision and deep learning improve quality control by identifying defects and anomalies in products with high precision, ensuring that only high-quality products reach the market. AI-powered demand forecasting and supply chain optimization enhance inventory management by predicting consumer demand more accurately and adjusting production schedules, accordingly, reducing waste and improving efficiency. Robotics integration in manufacturing processes increases automation, reduces human error, and enhances production speed and consistency.

2.2.7 Retail

The growth in the retail sector, and e-commerce platforms like Amazon, are sustained through features enabled by AI such as personalized recommendations, dynamic pricing, and customer experiences. Machine learning models analyze customer data, including browsing and purchase history, to provide personalized product recommendations and targeted marketing. It is not only large multinational companies that have benefited; local Indonesian e-commerce platforms like Tokopedia and Bukalapak and Singapore-based Shopee also use AI to suggest products based on a customer's past behavior, thereby increasing sales and customer satisfaction. AI-powered chatbots and virtual assistants offer 24/7 customer support, handling routine inquiries, and improving response times. These systems can assist with tasks such as order tracking, returns processing, and product information, freeing up human staff to focus on more complex issues.

Chapter 3
AI and Cybersecurity

As AI technologies become more advanced and pervasive, they present two primary areas of cybersecurity risk: attacks against AI, and attacks facilitated and enhanced by AI. Like any technology or system, AI is at risk of compromise or manipulation; however, there are unique aspects to how AI is developed and deployed that require special attention. For example, an attacker could subtly manipulate training data in a way that leads to generating misinformation or producing erroneous output that could negatively impact essential or critical systems. Conversely, AI can be a tool for the enhancement of existing tactics, techniques, and procedures (TTPs) in cyberattacks. That is, AI lowers the access barrier for cybercriminals, enabling individuals with minimal technical knowledge to launch sophisticated cyberattacks. For instance, AI-powered tools can automate the creation of malicious software and enhance social engineering attacks, making it easier for less skilled attackers to deploy ransomware or malware via highly convincing phishing messages.

The evolution of AI technology also blurs the line between synthetic media and human-generated content, complicating the detection of deepfakes. These highly realistic deepfakes pose significant threats, from disinformation campaigns to identity theft. Detection technologies struggle to keep up with the rapid advancements in AI-generated content, making these deepfakes more targeted and dangerous than ever before.

Despite these challenges, there is room for optimism. Cisco's Executive Vice President and Chief Product Officer Jeetu Patel has asserted, "It is a great time for tipping the scales in favor of the defenders."[17] This suggests that, while AI introduces new risks, it also offers unprecedented opportunities to enhance cybersecurity defenses.

It is important to also understand that AI-specific risks do not exist in isolation but rather within the context of traditional cybersecurity threats.

The first step in developing robust defenses is to fully understand how attacks occur and the associated risks. In this chapter, we have compiled a taxonomy of cybersecurity risks from a short literature review of existing frameworks from the Data Security Council of India (DSCI)[18] and US National Institute of Standards and Technology (NIST).[19]

3.1 Data Security Risks

This category encompasses risks related to the confidentiality, integrity, and privacy of data used to train and operate AI models. Attacks targeting these areas aim to manipulate, steal, or infer sensitive information about the training data, posing significant threats to the overall security and trustworthiness of AI systems.

3.1.1 Data Poisoning

Data poisoning involves attackers manipulating training data to introduce vulnerabilities or backdoors into an AI model. This malicious activity can significantly undermine the model's accuracy and reliability.

Example: An attacker makes changes to a dataset used to train a malware detection model. By carefully editing malware samples in the dataset, the attacker can cause the model to misclassify certain types of malwares as benign software. This misclassification can allow malicious software to bypass detection, leading to security breaches and potential damage to systems and data.

Plausible Real-World Scenario: In a hypothetical scenario, a healthcare AI system designed to diagnose diseases from medical images could be undermined if attackers manipulate the training data. By introducing images with subtle but impactful alterations, the AI model could be trained to misdiagnose certain conditions, potentially leading to incorrect treatment recommendations.

3.1.2 Data Extraction

Data extraction attacks involve attackers inferring or reconstructing sensitive information about the training data from the model's outputs or behavior. This type of attack can reveal confidential information, posing significant privacy or other risks relating to proprietary information.

Example: An attacker queries a language model with carefully crafted prompts designed to extract sensitive information embedded within the model's training data. For instance, the attacker might input prompts that lead the language model to reveal private details such as email addresses, phone numbers, or personal identifiers that were part of the original training data.

Plausible Real-World Scenario: In the context of a customer service chatbot trained on a vast dataset of customer interactions, an attacker could exploit the model to extract sensitive customer information. By using specific queries, the attacker might retrieve confidential details about past transactions or personal data, violating user privacy and potentially leading to identity theft or financial fraud.

3.1.3 Inference Attacks

Inference attacks allow attackers to determine whether a specific data point was part of the model's training dataset (membership inference) or infer global properties about the training data distribution (property inference). These attacks can compromise the confidentiality and integrity of the training data.

Example: An attacker analyzes the output of a machine learning model to infer whether a specific individual's data was used in the training dataset. This membership inference can reveal participation in the dataset, potentially exposing personal or sensitive information. Similarly, property inference allows attackers to deduce broader characteristics of the training data, such as the presence of certain demographic groups.

Plausible Real-World Scenario: Consider a machine learning model trained on financial transaction data to detect fraudulent activities. An attacker could exploit this model to infer whether a particular individual's transaction history was included in the training data. By carefully analyzing the model's responses to various inputs, the attacker might reveal sensitive financial behaviors or patterns, compromising the individual's privacy and potentially leading to targeted financial crimes.

3.2. Model Security Risks

Model security risks encompass threats that specifically target the AI model's architecture, parameters, or decision boundaries. These attacks aim to compromise the integrity and reliability of the model's outputs, potentially leading to harmful consequences. Understanding and mitigating these risks is crucial for ensuring the security and robustness of AI systems.

3.2.1 Evasion Attacks

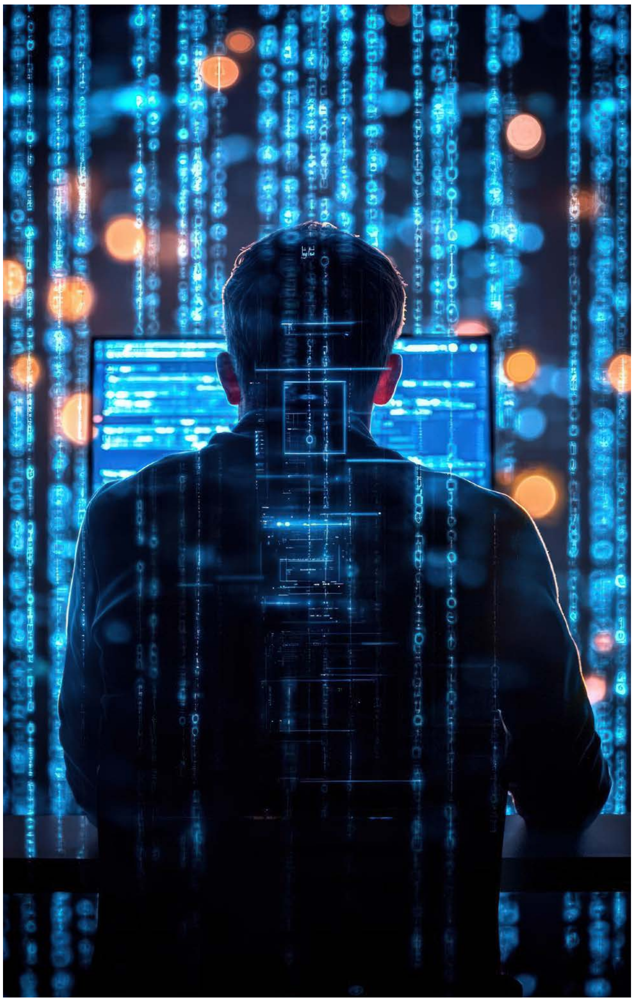
Evasion attacks involve attackers crafting adversarial examples designed to cause the model to misclassify or generate incorrect outputs. These subtle alterations can deceive even highly accurate models, leading to potentially dangerous outcomes.

Example: An attacker crafts minor, adversarial changes to an image, causing an autonomous vehicle's computer vision system to misclassify a stop sign as a speed limit sign. This misclassification could result in the vehicle failing to stop at an intersection, leading to accidents and endangering lives.

Plausible Real-World Scenario: In 2018, researchers demonstrated that by placing small stickers on stop signs, they could trick an AI-based traffic sign recognition system into misclassifying the signs. Such evasion attacks highlight the vulnerabilities in AI models used in critical applications like autonomous driving.[20]

3.2.2. Backdoor Attacks

Backdoor attacks involve attackers injecting hidden triggers into the model during its training phase. When these triggers are activated, they cause the model to make targeted misclassifications, often without altering the model's performance on regular inputs.



Example: An attacker poisons a facial recognition model during training by injecting images with a specific accessory, such as a pair of glasses. When individuals wear this accessory, the model misclassifies them and fails to identify them correctly. This can allow unauthorized individuals to bypass security systems.

Plausible Real-World Scenario: In a study, researchers found that by injecting poisoned training data into facial recognition systems, they could create backdoors that allowed them to bypass security measures simply by wearing a specific pair of glasses.[21] This technique could be used to evade security systems in high-stakes environments such as airports or secure facilities.

3.2.3 Model Poisoning

Model poisoning involves attackers manipulating the model's parameters or architecture to introduce vulnerabilities or degrade its performance. This is particularly concerning in collaborative learning environments such as federated learning.

Example: In a federated learning setting, where multiple participants contribute to training a global model, a malicious participant intentionally sends corrupted model updates to the central server. These poisoned updates degrade the global model's performance, potentially causing it to fail in critical applications.

Plausible Real-World Scenario: Federated learning is used in healthcare for collaborative training of predictive models across multiple hospitals without sharing patient data. If a malicious entity injects poisoned updates, the resulting model might provide inaccurate medical predictions, potentially harming patients.

3.2.4 Model Stealing

Model stealing occurs when attackers extract or replicate the model's architecture, parameters, or functionality through repeated query access. This can compromise proprietary models and intellectual property, enabling attackers to create unauthorized replicas.

Example: An attacker repeatedly queries an API-accessible machine learning model, analyzing the outputs to reconstruct a functionally equivalent model. This stolen model can then be used without authorization, potentially bypassing protections or creating competitive disadvantages.

Plausible Real-World Scenario: Researchers have demonstrated that by querying commercial machine learning APIs such as those offered by Google, Amazon, and Microsoft, they could effectively reconstruct the underlying models.[22] This kind of attack can lead to significant intellectual property theft and reduce the competitive advantage of AI service providers.

3.3. Infrastructure Risks

Infrastructure risks encompass threats related to the underlying infrastructure and tools used to develop, train, and deploy AI models. These attacks target the availability, performance, and supply chain integrity of AI systems, potentially leading to significant disruptions and vulnerabilities.

3.3.1 Denial of Service Attacks

Denial of service (DoS) attacks involve overwhelming an AI system with an excessive number of requests, rendering it unavailable or unresponsive to legitimate users. These attacks can disrupt critical services and degrade the overall user experience.

Example: An attacker floods an AI-powered image classification service with a high volume of requests. This deluge of traffic causes the service to become overloaded and unable to process legitimate user requests, resulting in service downtime and operational disruptions.

Real-World Scenario: A Distributed Denial of Service (DDoS) attack targeted the GitHub repository hosting platform. Although not an AI-specific attack, it highlighted how overwhelming requests can bring down critical infrastructure.[23] Similar tactics can be applied to AI systems, where the objective is to disrupt services by exhausting system resources.

3.3.2 Resource Exhaustion Attacks

Resource exhaustion attacks deplete the computational resources of an AI system, thereby degrading its performance or availability.

Attackers craft queries or inputs that exploit weaknesses in the system, leading to excessive consumption of computing resources and increased latency for legitimate requests.

Example: An attacker identifies a flaw in an AI system that allows them to submit computationally intensive queries. By repeatedly sending these queries, the attacker causes the AI system to consume an excessive amount of computing power, resulting in slow response times and degraded performance for legitimate users.

Real-World Scenario: In the context of cloud-based AI services, resource exhaustion attacks can lead to increased operational costs due to the excessive consumption of computational resources. For instance, an AI model deployed on a cloud platform could experience significant slowdowns and increased costs if targeted by such an attack.

3.3.3 Supply Chain Attacks

Supply chain attacks compromise the integrity of AI development tools, libraries, or platforms, introducing vulnerabilities that can be exploited by attackers. These attacks can have far-reaching consequences, as compromised components are often widely used across multiple AI systems and applications.

Example: An attacker compromises a widely used open-source machine learning framework, introducing a vulnerability that allows them to access sensitive data processed by AI systems built using this framework. This can lead to widespread data breaches and the potential exploitation of vulnerable AI systems.

Real-World Scenario: The SolarWinds supply chain attack highlighted the potential impact of such attacks on software infrastructure. A compromised update to SolarWinds' software led to widespread infiltration of government and corporate networks.[24] Similarly, an attack on AI development tools could compromise the security of numerous AI systems.

3.4 Application Risks

Application risks encompass the threats and challenges associated with the specific application or use case of an AI system. These risks can arise from the interaction between the AI model and its users, as well as from the broader ethical and societal implications of deploying AI in various contexts. Addressing these risks is essential for ensuring the responsible and ethical use of AI technologies.

3.4.1 Prompt Injection Attacks

Prompt injection attacks involve attackers manipulating the input prompts given to an AI system, causing it to generate unintended, harmful, or biased outputs. These attacks exploit the system's reliance on input data to influence its behavior in ways that can be malicious or misleading.

Example: An attacker crafts a misleading prompt that causes a language model to generate content promoting a specific political agenda, disguised as objective news. This can mislead users and spread disinformation, impacting public opinion and trust in information sources.

Plausible Real-World Scenario: On social media platforms, attackers are prone to prompt injection to spread propaganda or fake news. For instance, during election periods, manipulated prompts can lead AI systems to generate and disseminate false information, influencing voter behavior and undermining democratic processes.

3.4.2 Output Integrity Issues

Output integrity issues arise when an AI system generates inconsistent, unreliable, or biased outputs. These issues can lead to decision-making errors, reputational damage, or privacy leakage, compromising the trust and effectiveness of the AI system.

Example: An AI-powered content moderation system fails to consistently identify and flag hate speech, leading to the spread of harmful content on a social media platform. This inconsistency can result in the platform being criticized for not effectively managing harmful content, damaging its reputation and user trust.

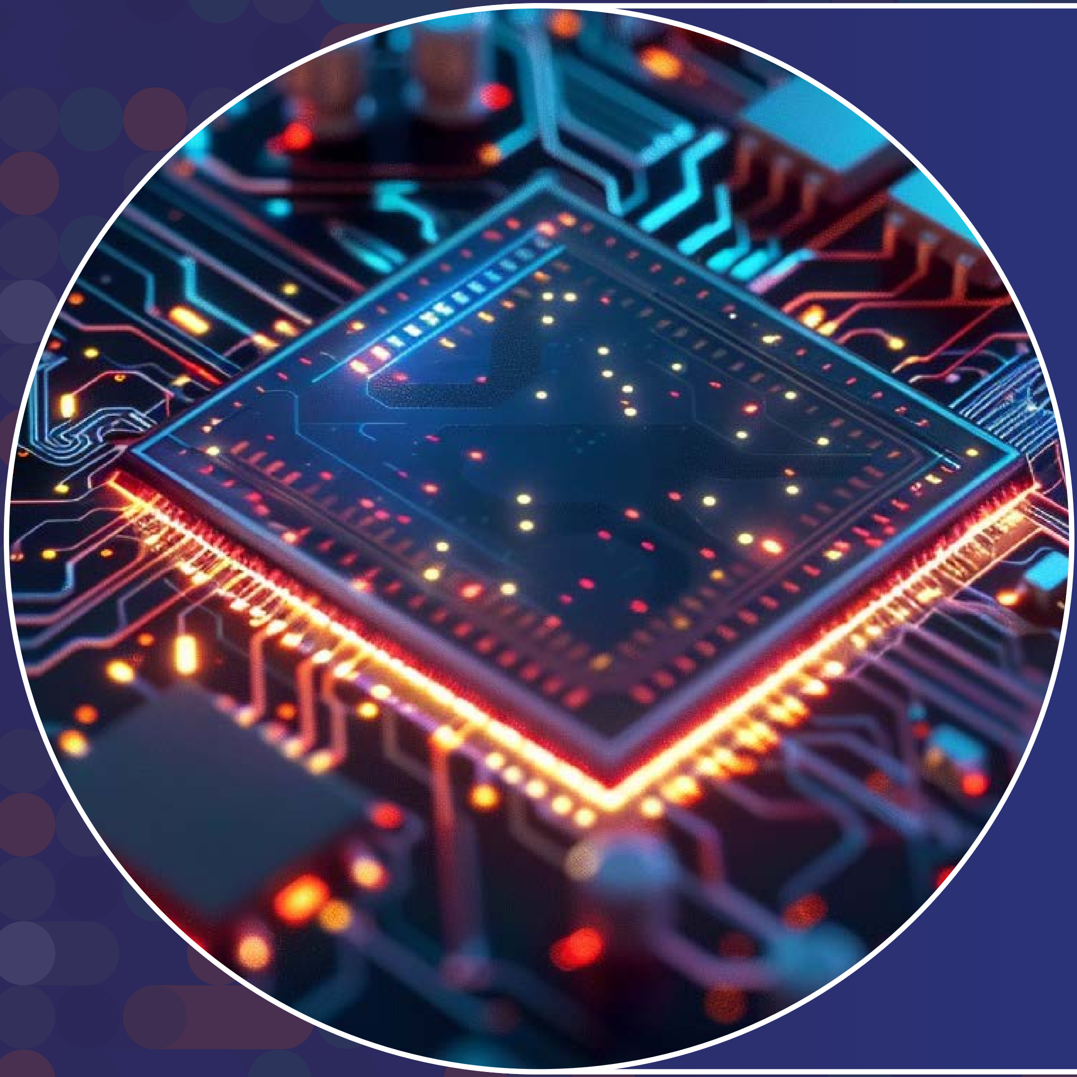
Plausible Real-World Scenario: Facebook and Twitter have faced criticism for their AI content moderation systems, which have sometimes failed to effectively identify and remove harmful content, such as hate speech and misinformation. These failures have resulted in significant public backlash and raised concerns about the reliability of AI in content moderation.

3.4.3 Ethical and Societal Risks

Ethical and societal risks involve the broader implications of AI systems on individuals, groups, or society. These risks can arise from the unintended consequences of AI use, leading to negative impacts such as reinforcing biases, violating privacy, or exacerbating social inequalities.

Example: An AI system used for predictive policing may learn and amplify biases present in historical crime data, leading to over-policing of certain communities and perpetuating social inequalities. This can result in increased surveillance and harassment of marginalized groups, undermining trust in law enforcement.

Plausible Real-World Scenario: The use of AI in predictive policing has been constantly scrutinized after reports showed that these systems disproportionately targeted minority communities. Studies indicate that historical biases in crime data led AI systems to unfairly focus on certain neighborhoods, raising ethical concerns about the fairness and impact of such technologies.[25]



Chapter 4:
Developing a Framework for Addressing AI Cybersecurity Risks

4.1 Approach

4.1.1 Building on existing laws and regulations

The Asia Pacific region, home to some of the world's most dynamic and rapidly evolving economies, is experiencing a surge in the adoption of AI. However, this technological advancement also exposes organizations to a complex landscape of cybersecurity threats and other risks related to the use of AI.

AI governance and oversight approaches should leverage existing laws and regulations to address emerging cybersecurity concerns. This strategy ensures regulatory consistency, avoids confusion from overlapping rules, and builds upon established best practices. Many current laws already cover key AI concerns such as data privacy, anti-discrimination, and intellectual property. By aligning AI cybersecurity measures with existing frameworks, organizations can streamline compliance, facilitate effective enforcement, and promote collaboration across sectors.

To this end, a step-by-step approach is essential in developing a framework for AI cybersecurity. This approach involves:

- Mapping existing regulations to identify and understand the regulatory landscape and pinpoint areas where current laws can be leveraged.
- Integrating best practices from existing regulations into the AI cybersecurity framework so that the framework benefits from established standards and methodologies.
- Tailoring existing regulations to address the unique risks associated with AI systems by updating regulatory requirements to cover AI-specific vulnerabilities, such as model evasion and data poisoning.
- Developing mechanisms for continuous monitoring and compliance checks so that AI systems can adhere to the updated regulatory requirements through regular audits, risk assessments, and reporting.

Key Regional Regulations

Several countries in the Asia Pacific region have developed legal frameworks that address AI cybersecurity risks. These include Singapore, Australia, Japan, and South Korea.

Singapore's Model AI Governance Framework, Personal Data Protection Act

Singapore's Model AI Governance Framework offers practical guidelines for organizations to manage AI deployment, focusing on issues like transparency, fairness, and accountability. This framework can guide the development of AI systems with built-in security measures, such as regular audits and risk assessments, to prevent and mitigate cybersecurity threats.[26] Additionally, Singapore's Personal Data Protection Act (PDPA) provides comprehensive guidelines on data protection, which is crucial for securing AI systems that rely on large datasets. The PDPA's principles of accountability and data minimization can be applied to AI systems to ensure that only necessary data is processed and that organizations are accountable for data protection.[27]

Australia's AI Ethics Framework, Security of Critical Infrastructure Act

Australia's AI Ethics Framework sets out principles to ensure the ethical deployment of AI, which includes considerations for security and privacy. The principle of "Protect Privacy and Security" from the AI Ethics Framework can be used to enforce robust data protection measures in AI systems, mitigating risks of data breaches and unauthorized access.[28] Further, Australia's Security of Critical Infrastructure

Act mandates stringent security practices for protecting critical infrastructure, including AI systems deployed in such sectors. Applying the Act's requirements for risk management and reporting to AI systems can enhance the resilience of AI applications in critical sectors like energy and finance.[29]

Japan's Basic Act on Cybersecurity, AI Utilization Guidelines

Japan's Basic Act on Cybersecurity establishes a legal basis for national cybersecurity efforts, emphasizing the protection of critical information infrastructure and the importance of cybersecurity in AI development. Leveraging the Act's emphasis on critical infrastructure protection can guide the implementation of robust cybersecurity measures in AI systems used in vital sectors.[30] Japan's AI Utilization Guidelines promote safe and secure AI usage, focusing on transparency, accountability, and user trust. The guidelines' focus on transparency can be applied to ensure that AI systems provide clear explanations of their processes and decisions, aiding in the detection and prevention of malicious activities.[31]

South Korea's Personal Information Protection Act (PIPA), National AI Strategy

South Korea's Personal Information Protection Act (PIPA) provides strong protections for personal data, essential for AI systems that handle sensitive information. The stringent data protection requirements of PIPA can be applied to AI systems to ensure that personal data is securely processed and stored, reducing the risk of data leaks and breaches.[32] South Korea's National AI Strategy outlines plans for AI development, emphasizing the need for robust cybersecurity measures to support AI innovation. The strategy's focus on cybersecurity can guide the integration of security-by-design principles in AI development, ensuring that AI systems are resilient against cyber threats from inception.[33]

ASEAN Guide on AI Governance and Ethics, ASEAN Framework on Digital Data Governance, Global Cross-Border Privacy Rules (CBPR) System

Promoting collaboration among countries fosters regional alignment in AI cybersecurity regulations, facilitating the development of regional standards and guidelines that support cross-border cooperation in addressing AI cybersecurity risks. Examples of frameworks include the ASEAN Guide on AI Governance and Ethics,[34] ASEAN Framework on Digital Data Governance and the Global Cross-Border Privacy Rules (CBPR) System. The ASEAN Framework on Digital Data Governance promotes data protection and privacy across ASEAN member states, providing a basis for alignment of AI cybersecurity measures. Implementing the principles of this framework in AI systems can ensure consistent data protection practices across the region, enhancing the security and trustworthiness of AI applications.[35] The CBPR system facilitates cross-border data flows while ensuring high standards of privacy protection. Aligning AI cybersecurity measures with the CBPR system can support secure data sharing and processing in AI applications, fostering innovation while protecting user privacy.[36]

4.1.2 Alignment with Internationally Recognized Standards

Beyond looking at regional frameworks, another method to strengthen cybersecurity policies is to look to internationally recognized standards for alignment. These standards provide comprehensive guidelines that cover a wide range of cybersecurity aspects, from technical specifications to ethical considerations, ensuring a holistic approach to AI governance.

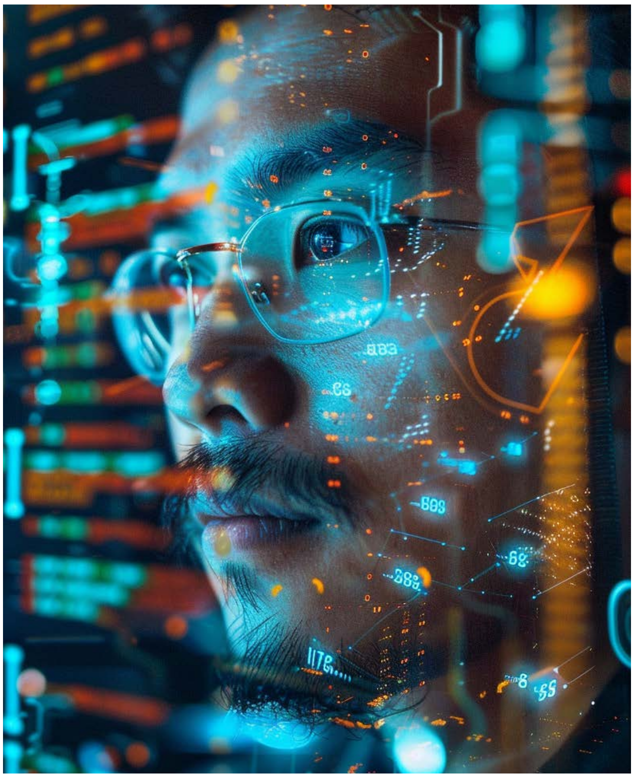
This alignment offers a structured approach to identifying, managing, and mitigating risks, which is essential for maintaining the integrity of AI technologies. Developed within the bounds of this structure, AI systems can seek to meet a global benchmark for security and reliability, fostering trust among users and stakeholders. For the Asia Pacific region, which comprises a diverse set of economies and regulatory environments, adhering to these standards can streamline compliance processes and reduce the regulatory burden on organizations, and enhance regional and global cooperation.

Technical Standards

Technical standards focus on the specific measures necessary to secure AI systems. These standards provide detailed guidelines on the implementation of security controls, ensuring that AI systems are designed and operated in a secure manner. Adherence to the latest security research and guidelines – such as those from NIST, International Organization for Standardization/ International Electrotechnical Commission (ISO/ IEC), Organization for the Advancement of Structured Information Standards (OASIS), and the Open Worldwide Application Security Project (OWASP) – is crucial for organizations to mitigate adversarial risks and align with advanced security practices. We cover some of key technical standard-making organizations below and their AI-related frameworks.

International Standards Organization (ISO) and International Electrotechnical Commission (IEC) Joint Technical Committee 1 SC 42

ISO and IEC often work together; in fact, IEC and ISO are the first international standards development organizations to set up an expert group to carry out standardization activities for AI. Sub Committee (SC) 42 is part of the joint technical committee ISO/IEC JTC 1. SC 42 considers the entire ecosystem in which AI systems are developed and deployed. It develops horizontal standards that provide a foundation upon which to create AI solutions for diverse industries. SC 42 works closely with IEC and ISO technical committees, where the focus is on sector-specific vertical standards.[37]



- The JTC1/SC42 are responsible for the following standards:
    - ISO/IEC 42001:2023 Information technology – Artificial intelligence – Management system[38]
    - ISO/IEC 23894:2023 Information technology – Artificial intelligence – Guidance on risk management[39]
    - ISO/IEC 23053:2022 Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)[40]
- Other relevant standards for cybersecurity also include ISO/IEC 27001 for information security management, which specifies the requirements for establishing, implementing, maintaining, and continually improving an information security management system (ISMS).[41] By aligning with this standard, organizations can ensure that their AI systems adhere to a structured framework for managing sensitive information and mitigating a broad set of security risks.

National Institute of Standards and Technology (NIST) ‘s AI Risk Management Framework and Cybersecurity Framework

The NIST AI Risk Management Framework (AI RMF) is a voluntary guideline designed to help organizations in the development and management of AI systems. It outlines a series of processes and activities that organizations can implement to manage risks throughout the AI lifecycle, addressing the involvement of various stakeholders.[42] Alongside the AI RMF, NIST has introduced a GitHub-hosted tool called the Playbook, which offers further recommendations on actions, references, and documentation that organizations can adopt.[43] As a living document, the AI RMF is expected to evolve in response to advancements in AI technology and shifts in the risk landscape.[44]

In response to the rise of Generative AI, NIST created a cross-sectoral profile and companion source to the AI RMF, the Generative Artificial Intelligence Profile. This profile seeks to define the risks that are novel or can be exacerbated by the use of Generative AI. It also provides a set of actions that organizations can adopt to “govern, map, measure, and manage” said risks.[45]

The NIST Cybersecurity Framework aims at helping organizations develop a comprehensive cybersecurity strategy that addresses the unique risks associated with AI systems. It provides a policy framework for computer security guidance, including detailed protocols for identifying, protecting, detecting, responding to, and recovering from cyber threats.[46] For example, the NIST framework’s guidelines on risk assessment can be used to identify potential vulnerabilities in AI models, such as adversarial attacks and data poisoning, and implement appropriate mitigation measures – guidance which is often lacking in most local legislation.

Other Internationally Recognized Standards

The Open Worldwide Application Security Project (OWASP)

The Open Worldwide Application Security Project (OWASP) is a non-profit foundation that aims to improve software security. The foundation offers an AI security and privacy guide to provide insights on “designing, creating, testing and procuring” AI systems that prioritizes privacy and security.[47]

MITRE Adversarial Threat Landscape for AI Systems (ATLAS)

MITRE ATLAS is a knowledge base of adversary tactics, techniques, and case studies of AI systems. It aims to inform government, industry, and academia of real-world threats to AI systems, enable threat assessments and read teaming, document adversary attacks on AI systems, and offers a framework for threat assessments and internal red teaming.[48]

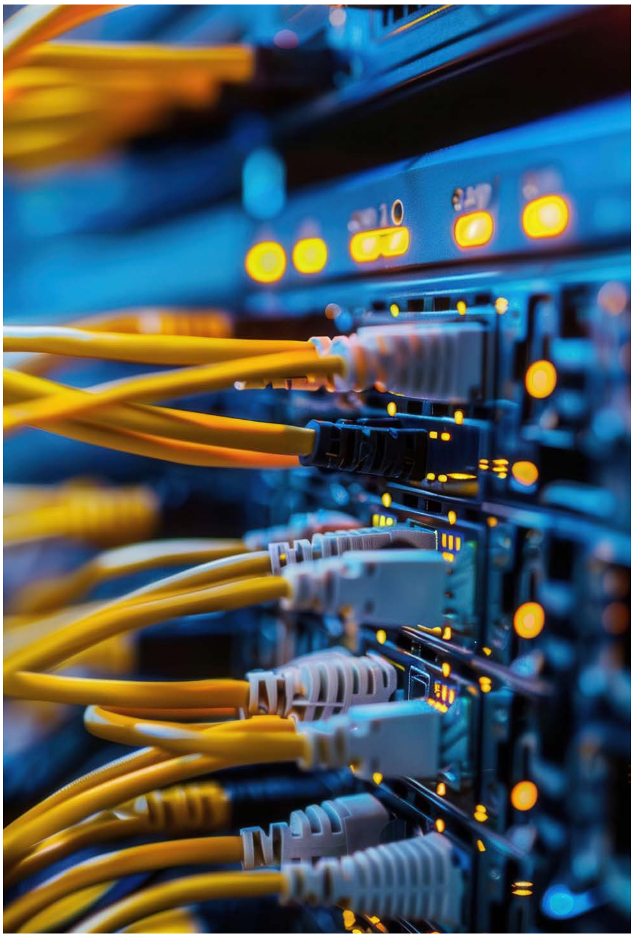
The Coalition for Secure AI (CoSAI)

The Coalition for Secure AI (CoSAI) is an international open-source consortium that seeks to address AI security issues through providing "open-source methodologies, standardized frameworks and tools".[49] CoSAI is further divided into various workstreams involving industry and academia in areas such as software supply chain security for AI systems, preparing defenders for a changing security landscape, and AI security governance.[50] The initiatives focus on integrating and leveraging AI securely across organizations through all phases of development and usage.

Alignment with internationally recognized standards and guidelines will lead to a risk-based approach to policymaking. By implementing established models such as the NIST Cybersecurity Framework, AI RMF ISO/IEC 27001, the AI RMF, COSO ERM, entities must proactively identify, assess, prioritize, and mitigate AI-related cybersecurity threats. The need for comprehensive risk management frameworks that align with internationally recognized standards that have been put through the rigor cannot be overstated. It ensures that resources are allocated efficiently, addressing the most critical threats first, and enhancing the overall security posture of the organizations.

Company Pledges to Cybersecurity and AI

Apart from regional regulations and international standards, over the past few years, various global companies have made significant pledges to enhance cybersecurity and ensure the responsible development and deployment of AI technologies. These commitments often emphasize the importance of security, transparency, and ethical considerations in AI:



Rome Call for AI Ethics (February 2020)

The landmark initiative known as the "Rome Call for AI Ethics" was spearheaded by the Vatican, bringing together leading technology companies such as Cisco, IBM, and Microsoft, alongside prominent academic institutions. This pledge aimed to promote an ethical approach to AI development, incorporating six core principles, one of which directly emphasized the need for security and privacy in AI systems. The Rome Call underscored the global recognition of the ethical and security challenges posed by AI and the collective responsibility to address these issues.[51]

US White House AI Commitments (July and September 2023)

The Biden-Harris Administration in the United States secured voluntary commitments from seven major AI companies, including Amazon, Anthropic, Google, Inflection, Meta, Microsoft, and OpenAI who pledged to prioritize the safe, secure, and transparent development of AI technologies.[52] In September 2023, additional commitments were obtained from companies such as Adobe, Cohere, IBM, NVIDIA, Palantir, Salesforce, Scale AI, and Stability.[53] These commitments represent a concerted effort by both the government and the private sector to manage the risks associated with AI and to ensure that AI technologies are developed with robust security measures in place.[54]

Bletchley Park AI Safety Summit (November 2023)

The Bletchley Park AI Safety Summit saw global technology companies and governments come together to pledge cooperation in addressing AI safety challenges. This Summit highlighted the recognition of the need for international collaboration in managing AI risks, with a particular focus on aligning efforts across different nations and industries to safeguard against potential AI-related threats.[55]

Seoul AI Safety Summit (May 2024)

At the Seoul AI Safety Summit, technology companies such as Amazon, Cisco, Google, Meta, and Microsoft, along with other prominent AI developers like OpenAI and Zhipu AI, signed a pledge to publish frameworks outlining how they will measure and mitigate risks associated with their “frontier” AI models.[56] This Summit emphasized the critical role of ongoing risk assessment and management in the safe deployment of advanced AI technologies.[57]

Thorne and All Tech Human Pledge (2024)

The Thorne and All Tech Human initiative gathered leading AI companies to pledge their commitment to protecting the safety of children online.[58] This pledge is part of a broader effort to address the societal impacts of AI and ensure that AI technologies are developed and deployed in a manner that prioritizes the well-being and security of vulnerable populations.[59]

These pledges reflect a growing recognition among both governments and corporations of the importance of cybersecurity and ethical considerations in AI development. By committing to these principles, these entities are taking proactive steps to address the potential risks associated with AI and to promote a safer, more secure digital future.

4.2 Key Components for Developing a Suitable AI Cybersecurity Framework

Developing a robust AI cybersecurity framework involves building on existing structures, aligning with global standards, and addressing the unique risks associated with AI systems.

In this endeavor, six key components are integral: Guidance and oversight, AI system lifecycle management, data governance and protection, model security and robustness, transparency and accountability, and incident response and recovery. These components are the building blocks for secure, reliable, and trustworthy AI systems.

4.2.1 Guidance and Oversight

There is a need for a cross-functional oversight committee comprising senior executives across an organization that would advise on AI practices and policies. It would also be the escalation and review point for high-risk use of AI.

Key Phase	Description
Leadership Committee	Provide leadership oversight through a Responsible AI Committee of senior executives across functions (e.g., sales, security, privacy, engineering, legal, human rights, government affairs, human resources).
Leadership oversight	Advise and monitor the organization on responsible AI practices and adoption of an AI governance framework.
Review use cases and incidents	Review sensitive or high-risk uses of AI being proposed and manage incident reports for bias or discrimination.

4.2.2 AI system lifecycle management

AI system lifecycle management involves overseeing the entire lifecycle of AI systems, from development and deployment to monitoring and decommissioning. Effective lifecycle regulation ensures that AI systems remain secure, reliable, and aligned with ethical standards throughout their operational lifespan.

Key Phase	Description
Development, Design and Training	Security and ethical considerations must be integrated by design. Adopting secure coding practices, conducting thorough threat modeling, and performing security testing are essential steps in this phase. This also extends to the selection, preparation, and integration of training assets. These should undergo the same thorough cyber security process to ensure they are free of attacks such as model poisoning.
Deployment	AI models must be securely integrated into production environments. Implementing strong access controls, ensuring secure configurations, and conducting vulnerability assessments are key measures to protect AI systems during this phase.
Monitoring, Maintenance and Model Draft Analysis	Continuous monitoring of AI systems is essential to detect and respond to emerging threats, and to ensure that the model is not drafting away from its design goals such that new attack vectors are being enabled. Implementing real-time monitoring, logging, and anomaly detection can help identify and mitigate security incidents promptly. Regular updates and patches are also necessary to address vulnerabilities.
Decommissioning	AI systems must be securely decommissioned when they reach the end of their lifecycle. This involves securely deleting data, deactivating models, and ensuring that no residual risks remain.

4.2.3 Model security and robustness

Regulating the security and robustness of AI models is crucial to protect against adversarial attacks and ensure reliable performance.

Key Phase	Description
Model Validation and Testing	Conducting thorough validation and testing of AI models can help identify and mitigate potential vulnerabilities. This includes testing for robustness, fairness, and bias.
Model Monitoring	Implementing automated monitoring and alerting systems can help maintain the integrity and performance of AI models.

4.2.4 Data governance and protection

Data governance and protection are fundamental to ensuring the security and privacy of data used in AI systems. Effective data governance involves implementing policies and procedures to manage data throughout its lifecycle, from collection and storage to processing and disposal.

Key Phase	Description
Data Privacy	Compliance with data protection regulations, such as the EU General Data Protection Regulation (GDPR) and the PDPA in Singapore, is critical. Implementing data anonymization, encryption, and access controls can help protect sensitive data.
Data Quality and Integrity	The accuracy and integrity of data are essential for the reliable performance of AI models. Implementing data validation, cleansing, and auditing processes can help maintain data quality.
Data Access Control	Implementing role-based access control (RBAC) and least privilege principles can help prevent unauthorized access to sensitive data. Regular access reviews and audits are necessary to ensure compliance with data governance policies.

4.2.5 Transparency and accountability

Transparency and accountability are critical to ensuring that AI systems operate ethically and can be trusted by users and stakeholders.

Key Phase	Description
Explainability	Implementing techniques for explainable AI (XAI) can help provide clear and understandable explanations of AI model decisions. This enhances transparency and helps build trust among users.
Auditability	Frameworks that ensure AI systems are auditable allow for independent verification of their performance and compliance with ethical standards. Implementing logging and audit trails can help achieve this.
Ethical Guidelines	Adopting ethical guidelines and principles, such as the OECD AI Principles and the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, can help ensure that AI systems are developed and deployed in a manner that respects human rights and promotes social well-being.

4.2.6 Incident response and recovery

Operational incident response and recovery mechanisms are vital components in mitigating the impact of security breaches and safeguarding the continuity of AI systems.

Key Phase	Description
Information Security Management Program	Steps should be taken to adopt comprehensive cybersecurity and supply chain security practices in line with internationally recognized standards (e.g., ISO, NIST)
Incident Detection and Response	Real-time monitoring and incident response systems can help detect and respond to security incidents promptly. Developing and regularly testing incident response plans is crucial for preparedness.
Disaster Recovery	Disaster recovery plans must be in place to help restore AI operations quickly in the event of a security breach or system failure. Regular testing and updating of these plans are necessary to ensure their effectiveness.
Continuous Improvement	Mandating a continuous improvement process involves learning from security incidents and updating policies, procedures, and technologies to enhance the resilience of AI systems.

Chapter 5:

Policy Recommendations – Managing AI Cybersecurity Risks and Enhancing AI Governance

5.1 Updating National Cybersecurity Strategy to Address AI Concerns

The integration of AI into various sectors has introduced new cybersecurity risks that require targeted policy responses. Effective management of these risks involves updating existing frameworks and introducing new regulations and practices tailored to the unique challenges of AI. This involves:

- **AI-Specific Threats:** Identifying and addressing AI-specific threats, such as data poisoning, model theft, and adversarial attacks, within national cybersecurity frameworks.
- **Cross-Sectoral Integration:** Ensuring that AI cybersecurity measures are integrated across all sectors, including healthcare, finance, transportation, and energy, to provide a cohesive defense against AI-related threats.
- **Public-Private Partnerships:** Encouraging collaboration between government agencies and private sector organizations to collaborate on policy development and share threat intelligence, best practices, and resources for combating AI cybersecurity threats.
- **Workforce Empowerment:** Catalyzing opportunities for new jobs and roles with AI accelerating the pace of change for the global workforce, providing workers with access to training programs, and enabling businesses to connect with skilled workers.

5.2 Establishing AI-Specific Cybersecurity Guidelines

Developing and implementing AI-specific cybersecurity guidelines can provide a clear framework to follow, ensuring consistent and effective protection across different sectors.

These guidelines should include:

- **Appropriate Oversight:** Instituting risk-based oversight of AI development and deployment focusing on well-defined, high-risk use cases and their associated impact.
- **AI Risk Assessment Protocols:** Establishing protocols based on internationally recognized, best practices for assessing the cybersecurity risks of AI systems throughout their lifecycle, from development to deployment and decommissioning.
- **Best Practices for Data Protection:** Defining best practices for data governance, including data integrity, privacy, and access control, to prevent data-related attacks on AI systems.
- **Model Risk-Based Measures:** Recommending security measures to protect AI models, such as adversarial training, robustness testing, and secure model update mechanisms.
- **Aligning to Internationally recognized Standards:** AI Cybersecurity Frameworks grounded in internationally recognized standards, but adaptive to emerging challenges, will be crucial in safeguarding against cybersecurity threats, thereby promoting secure and responsible AI innovation.

5.3 Investing in AI Cybersecurity Research and Development

Investing in research and development (R&D) focused on AI cybersecurity is crucial for staying ahead of emerging threats and developing innovative solutions. Key areas for investment should include:

- **Advanced Threat Detection:** Encouraging the development of advanced techniques for detecting and mitigating AI-specific threats.
- **Secure AI Frameworks:** Creating secure AI development frameworks and tools that incorporate security by design principles and facilitate the building of robust, trustworthy AI systems.
- **Collaboration with Academia:** Fostering partnerships between government agencies, industry, and academic institutions to drive research and innovation in AI cybersecurity.

5.4 Fostering International Cooperation and Coordination

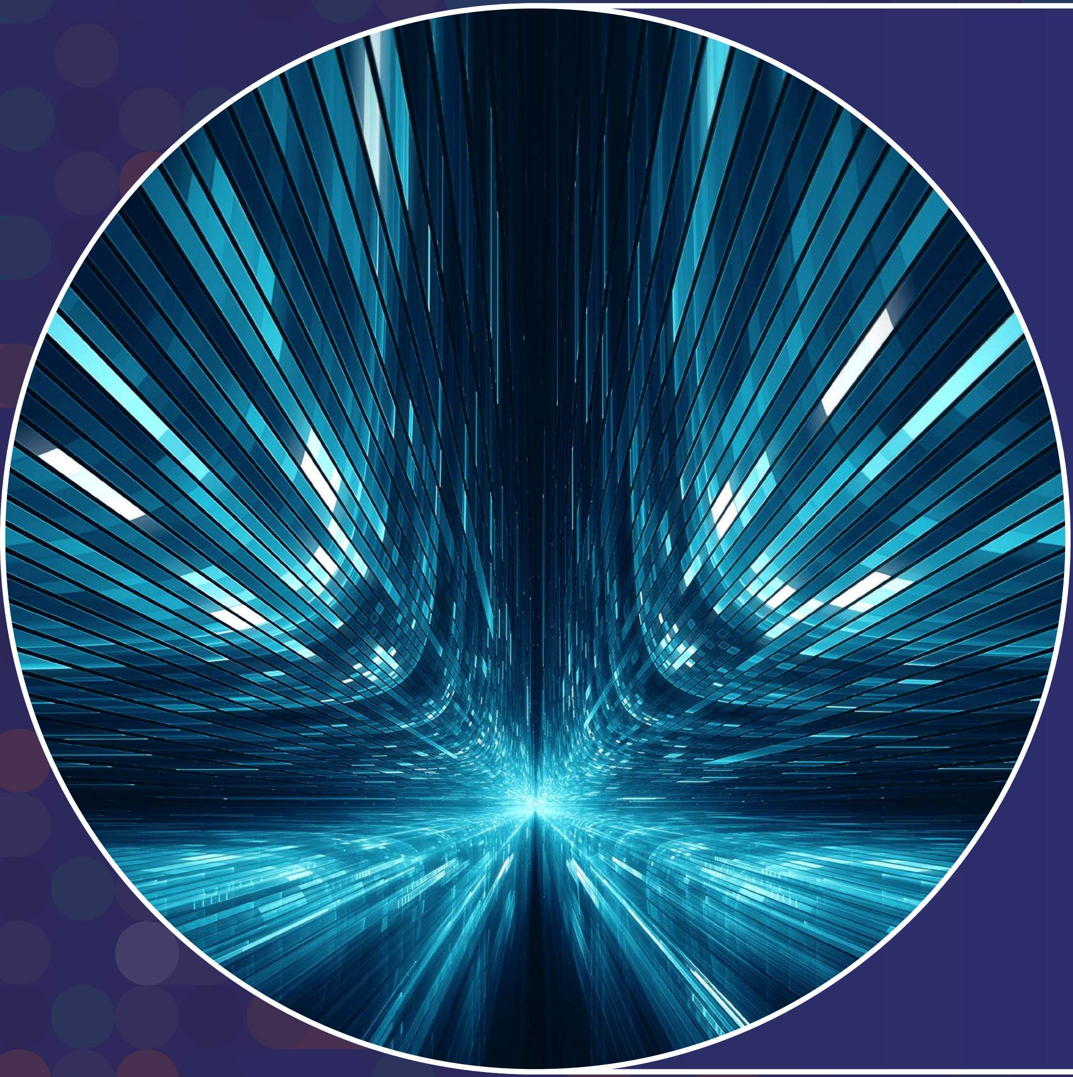
AI cybersecurity challenges are global in nature and require coordinated international efforts to address effectively. Policymakers should prioritize:

- **Alignment of Standards:** Working towards the alignment of AI cybersecurity standards and regulations across the region and with global partners to ensure a unified approach to managing AI risks.
- **Information Sharing:** Establishing cross-border mechanisms for sharing threat intelligence, best practices, and lessons learnt from AI cybersecurity incidents.
- **Joint Initiatives:** Participating in joint initiatives and collaborative projects with international partners to develop and implement AI cybersecurity solutions.

5.5 Promoting AI Literacy and Workforce Development

Building a skilled workforce and promoting AI literacy are essential for enhancing the cybersecurity of AI systems. This involves:

- **Educational Programs:** Developing educational programs and curricula that focus on AI cybersecurity, ensuring that the next generation of professionals is equipped with the necessary skills and knowledge.
- **Professional Training:** Offering training and certification programs for current professionals to keep them updated on the latest AI cybersecurity threats, techniques, and best practices.
- **Public Awareness Campaigns:** Conducting public awareness campaigns to educate citizens about AI cybersecurity risks and how to protect themselves from AI-related threats.



Chapter 6
Conclusion

This report highlights the critical importance of ensuring AI safety and cybersecurity, emphasizing that as AI technologies become more integrated across various sectors, the associated risks must be effectively managed.

Developing a comprehensive AI cybersecurity framework requires building on existing policies, aligning with internationally recognized standards and practices, and addressing the unique risks associated with AI systems. Six key components are integral to this framework: guidance and oversight, AI system lifecycle management, data governance and protection, model security and robustness, transparency and accountability, and incident response and recovery.

To mitigate AI-related cybersecurity risks and strengthen AI governance, national cybersecurity strategies should be updated to address AI-specific threats, integrating measures across sectors, and fostering public-private partnerships. Further, AI-specific guidelines, including risk-based oversight and data protection best practices, should be established. Invest in AI cybersecurity research and development is also necessary to advance threat detection and secure frameworks. International cooperation is needed for aligning standards and sharing threat intelligence. Additionally, promoting AI literacy and workforce development through educational, professional training, and public awareness campaigns is vital to equipping individuals with the skills to mitigate AI-related risks.

Addressing the complexity of AI demands a coordinated approach across governments, industries, and technology companies. It is crucial to embed the right principles into the design and deployment of AI technologies to ensure responsible innovation. Shared standards and continuous collaboration will be essential for navigating the evolving landscape of AI and cybersecurity, so that the benefits of AI can be realized while its risks are effectively mitigated.

Endnotes

1 Cisco's AI Readiness Index (https://www.cisco.com/c/m/en_us/solutions/ai/readiness-index.html) that surveyed over 8,000 global companies found 84% of companies think AI would have a very significant or significant impact on their business and 97% of companies say the urgency to deploy AI-powered technologies increased. There is a profound gap between accelerating pace of AI development and how prepared organizations are in adopting it.

2 https://www.ncsc.gov.uk/guidance/ai-and-cyber-security-what-you-need-to-know

3 https://aaai.org/

4 https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021PC0206

5 https://theconversation.com/a-brief-history-of-ai-how-we-got-here-and-where-we-are-going-233482

6 https://www.forbes.com/sites/bernardmarr/2016/02/19/a-short-history-of-machine-learning-every-manager-should-read/

7 https://news.mit.edu/2023/explained-generative-ai-1109

8 https://aisel.aisnet.org/pacis2021/44

9 https://www.researchgate.net/profile/Mani-Madhukar/publication/316202671_IBM's_Watson_Analytics_for_Health_Care/links/5e2bed2d92851c3aadd7d440/IBMs-Watson-Analytics-for-HealthCare.pdf?_tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6InB1YmxpY2F0aW9uIiwicGFnZSI6InB1YmxpY2F0aW9uIn19

10 https://www.cisco.com/c/m/en_us/solutions/predictive-networks/index.html

11 https://asean.org/wp-content/uploads/2021/01/fa-220416_DM2025_email.pdf

12 https://www.kasparov.com/timeline-event/deep-blue/

13 https://www.smartnation.gov.sg/nais/

14 https://www.technologyreview.com/2023/12/29/1084699/machine-learning-earthquake-prediction-ai-artificial-intelligence/

15 https://www.bsp.gov.ph/SitePages/InclusiveFinance/InclusiveFinance.aspx

16 https://newsroom.ibm.com/2021-02-04-EGAT-Adopts-IBM-AI-to-Help-Improve-Efficiency-Across-Thailands-Major-Power-Plants

17 https://www.weforum.org/agenda/2024/02/ai-cybersecurity-how-to-navigate-the-risks-and-opportunities/

18 https://www.dsci.in/resource/content/mitigating-security-privacy-risks-guide-enterprise-use-generative-ai

19 https://csrc.nist.gov/pubs/ai/100/2/e2023/final

20 https://ieeexplore.ieee.org/document/8578273

21 https://arxiv.org/abs/1708.06733

22 https://dl.acm.org/doi/10.5555/3241094.3241142

23 https://www.a10networks.com/blog/5-most-famous-ddos-attacks/

24 https://www.techtarget.com/whatis/feature/SolarWinds-hack-explained-Everything-you-need-to-know

25 https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3333423

26 https://www.pdpc.gov.sg/help-and-resources/2020/01/second-edition-of-model-artificial-intelligence-governance-framework

27 https://sso.agc.gov.sg/Acts-Supp/40-2020

28 https://www.industry.gov.au/publications/australias-artificial-intelligence-ethics-framework

29 https://www.cisc.gov.au/legislation-regulation-and-compliance/soci-act-2018

30 https://www.japaneselawtranslation.go.jp/en/laws/view/3677/en

31 https://www.soumu.go.jp/main_content/000658284.pdf

32 https://elaw.klri.re.kr/eng_mobile/viewer.do?hseq=62389&type=part&key=4

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

34 https://asean.org/book/asean-guide-on-ai-governance-and-ethics/

35 https://asean.org/wp-content/uploads/2012/05/6B-ASEAN-Framework-on-Digital-Data-Governance_Endorsedv1.pdf

36 https://www.globalcbpr.org/wp-content/uploads/Global-CBPR-Framework-2023.pdf

37 https://www.iec.ch/ords/f?p=103%3A7%3A704219401578297%3A%3A%3A%3AFSP_ORG_ID%3A21538

38 https://www.iso.org/standard/81230.html

39 https://www.iso.org/standard/77304.html

40 https://www.iso.org/standard/74438.html

41 ISO/IEC 27001 https://www.iso.org/standard/27001

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

43 https://www.brookings.edu/articles/nists-ai-risk-management-framework-plants-a-flag-in-the-ai-debate/

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

45 https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf

46 https://www.nist.gov/cyberframework

47 https://owasp.org/www-project-ai-security-and-privacy-guide/

48 https://atlas.mitre.org/pdf-files/MITRE_ATLAS_Fact_Sheet.pdf

49 https://www.oasis-open.org/2024/07/18/introducing-cosai/

50 https://www.coalitionforsecureai.org/about/

51 https://www.romecall.org/the-call/

52 https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/

53 https://www.whitehouse.gov/briefing-room/statements-releases/2023/09/12/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-eight-additional-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/

54 https://www.whitehouse.gov/briefing-room/statements-releases/2023/09/12/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-eight-additional-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai

55 https://www.aisafetysummit.gov.uk

56 https://apnews.com/article/south-korea-seoul-ai-summit-uk-2cc2b297872d860edc60545d5a5cf598

57 https://aiseoulsummit.kr/press/?mod=document&uid=43

58 https://www.thorn.org/blog/generative-ai-principles/

59 https://www.thorn.org/blog/generative-ai-principles

COALITION FOR CYBERSECURITY IN ASIA-PACIFIC

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.

https://ccapac.asia