ประชุมสัมมนาโครงการสร้างเครือข่ายสำนักวิทยบริการและเทคโนโลยีสารสนเทศ
มหาวิทยาลัยเทคโนโลยีราชมงคล ครั้งที่ 9 (ARIT Net #9)

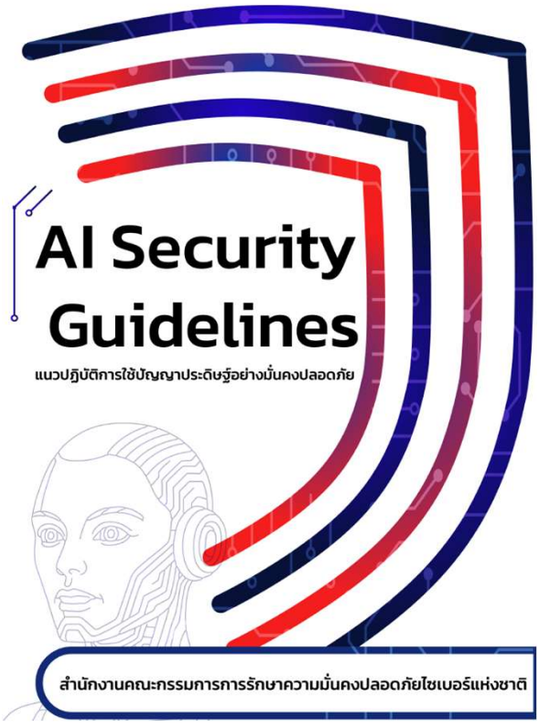
AI Security Guidelines จากแนวคิดสู่การลงมือปฏิบัติ ด้วยหลักจริยธรรมและปลอดภัยอย่างยั่งยืน

รองศาสตราจารย์ ดร.พงษ์พิสิฐ วุฒิดิษฐโชติ

Thailand Cyber Security & Privacy Officer (CSPO)

Huawei Technologies (Thailand) Co.,Ltd.

pongpisit.wuttidittachotti@huawei.com



26 กุมภาพันธ์ 2569

Version 1.0 ปรับปรุงล่าสุด 26 กุมภาพันธ์ 2569 เวลา 09.00 น.

- หัวหน้าภาควิชาการบริหารเครือข่ายดิจิทัลและความมั่นคงปลอดภัยสารสนเทศ คณะเทคโนโลยีสารสนเทศและนวัตกรรมดิจิทัล มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ ตั้งแต่ กรกฎาคม 2559- 18 สิงหาคม 2567
- Digital Talent Award of the Year 2022, Digital Community of the year จากผลงานโครงการแอปพลิเคชั่น เพื่อสร้างเกษตรฟาร์มใหญ่บริหารจัดการข้อมูลผลิตผลและการตลาดสำหรับใหญ่เครือข่ายเกษตรกร กรณีศึกษา สหกรณ์การเกษตรเชียงกลาง จำกัด จังหวัดเลย
- อาจารย์ดีเด่นแห่งชาติ สาขารับใช้สังคม ปี 2565,
  ที่ประชุมประธานสภาอาจารย์มหาวิทยาลัยแห่งประเทศไทย (ปอมท.), https://youtu.be/NH7h8MzJ0dQ

SCAN ME

- ที่ปรึกษาโครงการจัดทำแผนตามพระราชบัญญัติการรักษาความมันคงมั่นคงปลอดภัยไซเบอร์ เพื่อสนับสนุนการดำเนินงานของสำนักงานคณะกรรมการการรักษาความมันคงมั่นคงปลอดภัยไซเบอร์แห่งชาติ (สกมช.) ปี 2563-2564
  - นโยบายและแผนปฏิบัติการว่าด้วยการรักษาความมั่นคงปลอดภัยไซเบอร์ พ.ศ. 2565-2570
  - ประมวลแนวทางปฏิบัติและกรอบมาตรฐานด้านการรักษาความมั่นคงปลอดภัยไซเบอร์ สำหรับหน่วยงานของรัฐและหน่วยงานโครงสร้างพื้นฐานสำคัญทางสารสนเทศ พ.ศ. 2564
- ประกาศนียบัตร (Certificate) นานาชาติด้านมาตรฐาน ความมั่นคงปลอดภัยไซเบอร์ ความเป็นส่วนตัว และการกำกับดูแล มากกว่า 100 รายการ เช่น CISA, CISM, CISSP, FIP
- ได้รับการรับรองเป็นผู้ปฏิบัติวิชาชีพวิศวกรรม (จากสภาวิศวกร) ระดับวิศวกรวิชาชีพ สาขา วิศวกรรมคอมพิวเตอร์ คนแรกของประเทศไทย เลขทะเบียนใบรับรอง วคพ.1-001

HUAWEI

Certificate

CISSP® CCSP™ CSSLP® HCISPP® CGRC™ CC℠

FIP
IAPP Fellow of Information Privacy

Certified Information Privacy Professional/Europe CIPP E iapp
Certified Information Privacy Manager CIPM iapp
Certified Information Privacy Technologist CIPT iapp

PCI Security Standards Council PCI Professional (PCIP) Certification

PCI Security Standards Council™
PAYMENT APPLICATION QUALIFIED SECURITY ASSESSOR

CISA CISM CRISC CGEIT CDPSE

CSX CYBERSECURITY NEXUS FUNDAMENTALS

CompTIA CASP+ CERTIFIED·CE
CompTIA Security+ CERTIFIED·CE
CompTIA CTT+ CERTIFIED
CompTIA Cloud+ CERTIFIED·CE

CompTIA Storage+™ POWERED BY SNIA CERTIFIED

CompTIA A+ CERTIFIED·CE
CompTIA Project+ CERTIFIED
CompTIA Server+ CERTIFIED

APMG INTERNATIONAL COBIT® 5 AN ISACA® FRAMEWORK FOUNDATION

COBIT 2019 Foundation An ISACA® Certificate

COBIT 2019 Design & Implementation An ISACA® Certificate

BUREAU VERITAS 1828 CEPAS DPO

aws certified Solutions Architect Associate

aws certified Cloud Practitioner

C|EH™ Certified Ethical Hacker

PECB Certified Trainer

CCSK™ Certificate of Cloud Security Knowledge

CCAK™ Certificate of Cloud Auditing Knowledge A Cloud Security Alliance® and ISACA® Credential

C|EI™ Certified EC-Council Instructor

EC-Council E|C|I|H CERTIFIED

PECB ISO/IEC 42001 SENIOR LEAD IMPLEMENTER

PECB ISO/IEC 27001 SENIOR LEAD AUDITOR

PECB ISO/IEC 27001 SENIOR LEAD IMPLEMENTER

PECB ISO/IEC 27701 SENIOR LEAD IMPLEMENTER

PECB ISO/IEC 27701 SENIOR LEAD AUDITOR

PECB ISO/IEC 27005 SENIOR LEAD RISK MANAGER

PECB ISO/IEC 38500 SENIOR LEAD IT CORPORATE GOVERNANCE MANAGER

PECB ISO/IEC 27035 LEAD INCIDENT MANAGER

PECB ISO/IEC 20000 SENIOR LEAD AUDITOR

PECB DATA PROTECTION OFFICER

PECB SENIOR LEAD CLOUD SECURITY MANAGER

- ติดอันดับหนังสือขายดีหลายสัปดาห์ ช่วงเมษายน-กรกฎาคม 2564
- หนังสือที่ทำให้เรื่องกฎหมาย PDPA และ Cybersecurity เข้าถึงประชาชนได้แบบเข้าใจง่ายเล่มแรก ๆ ของประเทศ
- ได้รับการยอมรับและตอบรับที่ดีจากผู้อ่านจำนวนมาก
- มีจำหน่ายตามร้านหนังสือชั้นนำทั่วประเทศ และร้านออนไลน์

HUAWEI

ISACA Educational Excellence Award

Recognizes an inspiring educator or academic at the primary, secondary or tertiary educational level or an educational institution with inspiring initiatives who has empowered students to pursue careers advancing technology.

Pongpisit Wuttidittachotti, CISA, CISM, CRISC, CGEIT, CDPSE, CCAK, COBIT2019, CISSP, CGRC

"For over two decades of profound contributions to GRC and cybersecurity education, empowering future leaders and strengthening enduring collaboration between academia and industry."

Dr. Pongpisit Wuttidittachotti's perspective positions knowledge as a catalyst for personal growth, national resilience, and societal strength. Rising from a rural orphaned childhood, he pursued learning with discipline, earning over 100 certifications, including key ISACA credentials, and applying them to national priorities. For over two decades, he advanced GRC and cybersecurity education, mentored thousands, shaped national frameworks, expanded high-quality learning access for millions, empowered ethical leaders, strengthened sustainable digital trust, deepened cross-sector collaboration, and is shaping an enduring legacy of impact.

2026 ISACA Educational Excellence Award

รางวัลเชิดชูเกียรติระดับสากลที่มอบให้แก่นักวิชาการผู้สร้างคุณูปการโดดเด่นด้านการศึกษาในสาขา GRC และความมั่นคงปลอดภัยไซเบอร์

โดยมีการพิจารณาคัดเลือกผู้ทรงคุณวุฒิเพียงหนึ่งท่านต่อปีจากทั่วโลก

HUAWEI

Case Study: “Master Key” Vulnerability in DJI Romo

เหตุการณ์สำคัญ (Incident Overview)

- นักวิจัยพยายามใช้ AI ช่วยวิเคราะห์ API ควบคุมหุ่นยนต์ดูดฝุ่น
- พบช่องโหว่ **Zero Device Verification** ใน Cloud Authentication
- Auth Token เดียว ควบคุมอุปกรณ์ได้ **7,000+ เครื่อง / 24 ประเทศ**
- ผู้ผลิตออกแพตช์แก้ไขต้น **ก.พ. 2026**

ลักษณะช่องโหว่ (Technical Impact)

- ❌ ไม่มีการผูก Token กับอุปกรณ์ (No Identity Binding)
- 🔓 เข้าถึงกล้อง/ไมโครโฟนแบบ Real-time + แผนผังบ้าน
- 🎮 สั่งการอุปกรณ์จากระยะไกลได้ทั่วโลก (Remote Hijacking)

บทเรียนเชิงกลยุทธ์ (Key Takeaways)

- **Architecture > Device**: ความปลอดภัยพังที่ Logic ของ Cloud ไม่ใช่ตัวอุปกรณ์
- **AI as Force Multiplier**: AI ทำให้การ Reverse Engineering เร็วและง่ายขึ้น
- **Privacy Risk สูงมาก**: Smart Home + Sensor = High Impact หาก Access Control ล้มเหลว

Gartner เตือนองค์กรระงับการใช้ AI Browsers ชั่วคราว

1. ความเข้าใจพื้นฐาน: AI Browser คืออะไร?

- **นิยาม:** เบราว์เซอร์ที่ฝัง AI Agents (เช่น Perplexity Comet, ChatGPT Atlas) ซึ่งทำงานได้อัตโนมัติมากกว่าการค้นหาข้อมูลแบบเดิม
- **ความสามารถ:** สรุปเนื้อหาเว็บ, ร่างอีเมล, นำทางและทำธุรกรรมบนเว็บได้เองโดยที่มนุษย์ไม่ต้องสั่งการทุกขั้นตอน

2. 4 ความเสี่ยงวิกฤต (Systemic Risks)

- **ข้อมูลรั่วไหลสู่ภายนอก (Data Leakage):** การประมวลผลเกิดขึ้นบน Cloud ของผู้ให้บริการ ข้อมูลความลับบริษัทอาจถูกส่งออกไปโดยไม่รู้ตัว
- **การทำงานผิดพลาดของ AI (Execution Errors):** AI อาจกรอกข้อมูลผิด, จองบริการผิดพลาด หรือเข้าเรียนคอร์สอบรมแทนพนักงานโดยไม่มีการเรียนรู้จริง
- **ช่องโหว่ด้านความปลอดภัย (Security Flaws):** เทคโนโลยีใหม่มักมีบั๊ก (เช่น กรณี ChatGPT Atlas พบช่องโหว่เข้าถึงบัญชีผู้ใช้หลังเปิดตัวไม่กี่วัน)
- **เน้นความสะดวกมากกว่าความปลอดภัย:** การตั้งค่าเริ่มต้นมักเก็บข้อมูลผู้ใช้ไว้ระยะยาวเพื่อนำไปฝึกฝน AI ต่อ

การ์ทเนอร์ แนะองค์กรระงับการใช้ AI Browser

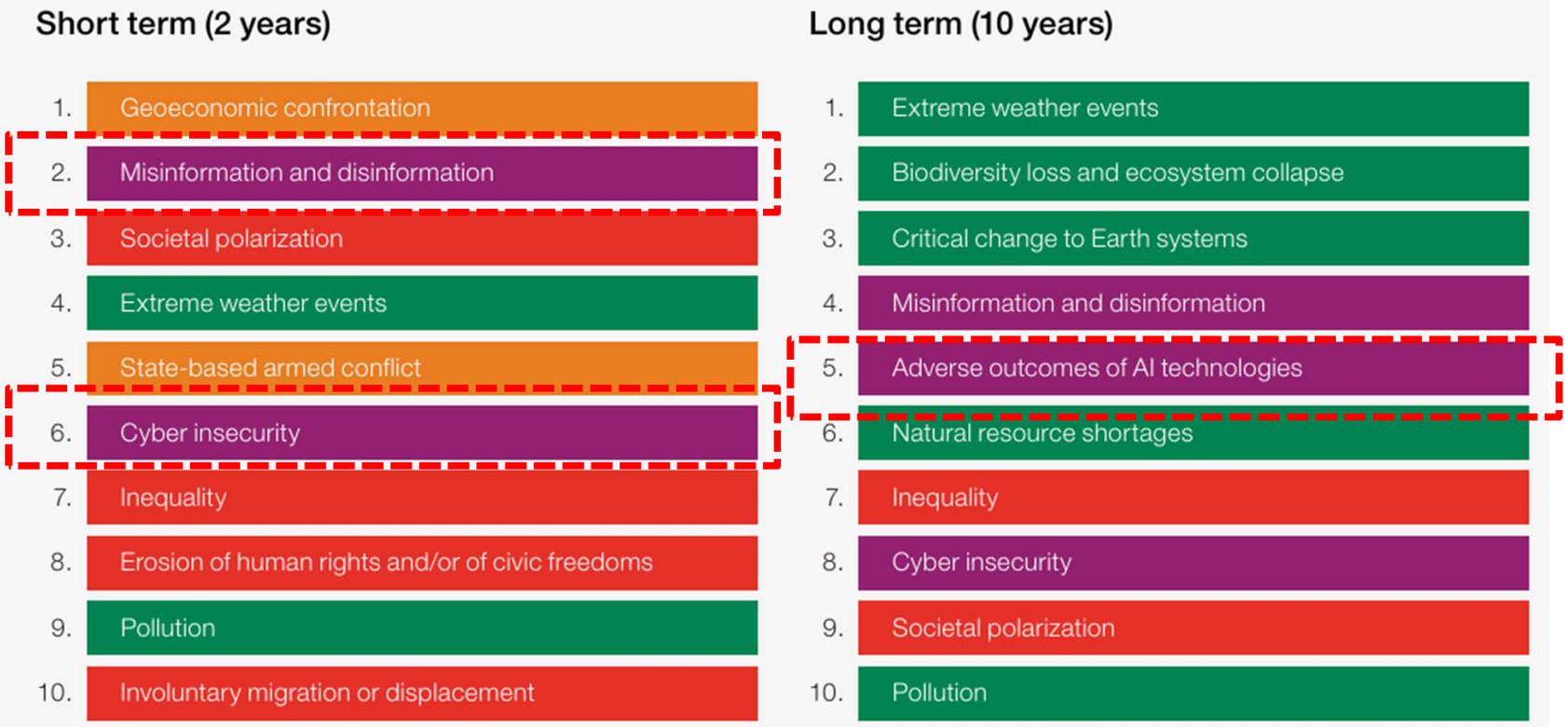
การเงินธนาคาร
20 กุมภาพันธ์ 2026 11:39 น.

https://moneyandbanking.co.th/2026/226591/

FIGURE 3

Global risks ranked by severity, short term (2 years) and long term (10 years)

"Please estimate the likely impact (severity) of the following risks over a 2-year and 10-year period."



Source

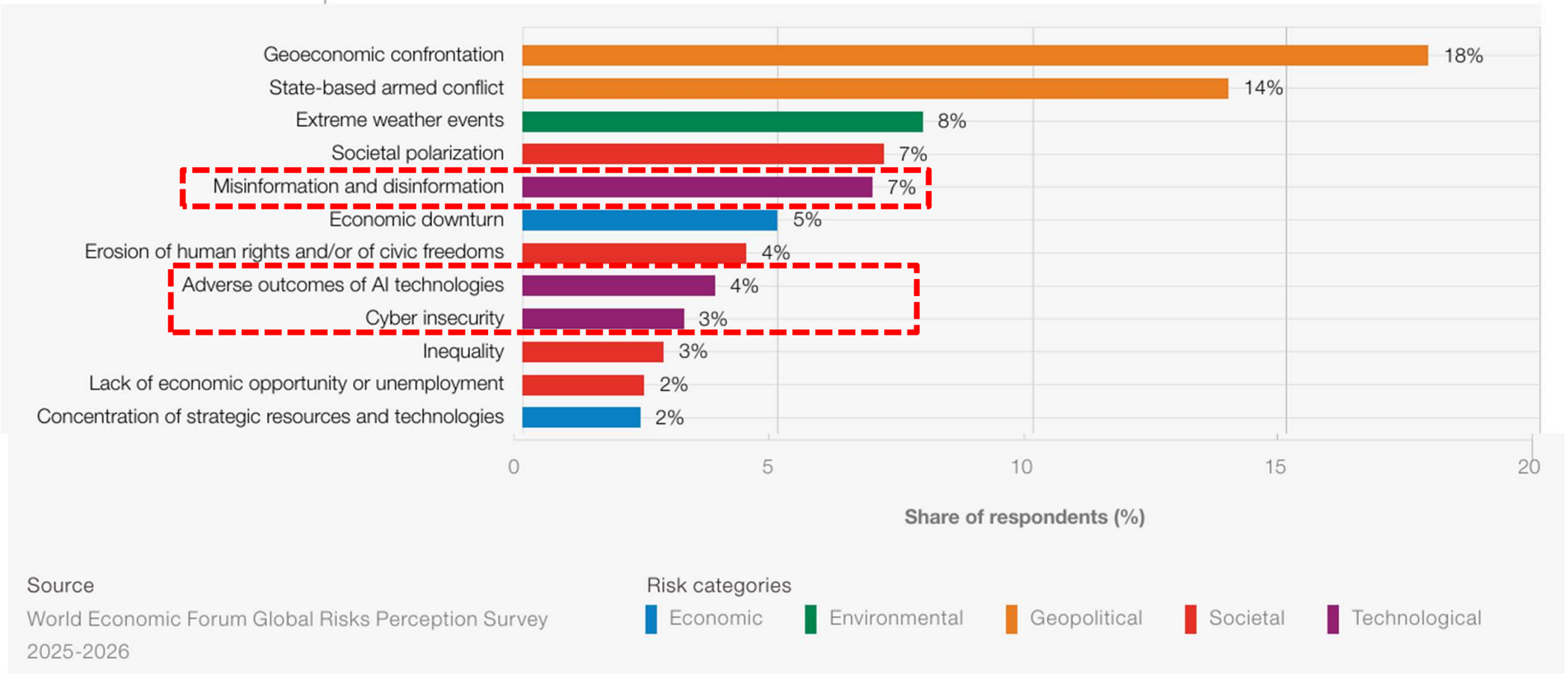
World Economic Forum Global Risks Perception Survey 2025-2026

Risk categories

Economic | Environmental | Geopolitical | Societal | Technological

FIGURE 2 **Current Global Risk Landscape**

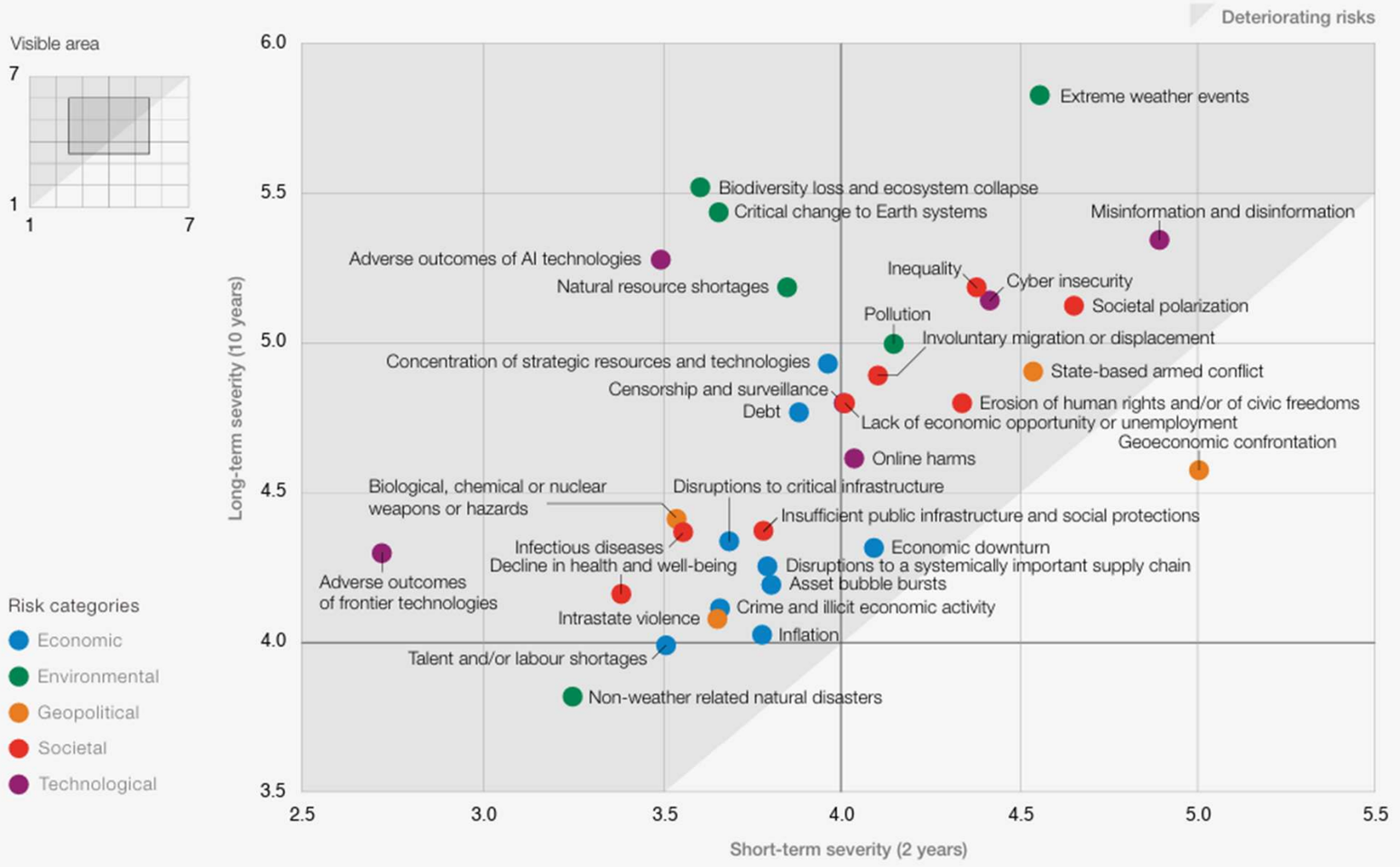
"Please select one risk that you believe is most likely to present a material crisis on a global scale in 2026."



Source

World Economic Forum Global Risks Perception Survey 2025-2026

FIGURE 7 **Relative severity of global risks, short term (2 years) and long term (10 years)**



Source
World Economic Forum Global Risks Perception Survey 2025-2026

Note
Severity was assessed on a 1-7 Likert scale [1 = Low severity, 7 = High severity].

International AI Safety Report 2026

ภาพรวมความเสี่ยงของปัญญาประดิษฐ์ (Overview of AI Risks)

รายงานฉบับนี้จัดกลุ่มความเสี่ยงของ AI ออกเป็น **3 ระดับหลัก** ได้แก่

1.การใช้งานโดยเจตนาร้าย

2.ความผิดพลาดหรือการทำงานผิดปกติ

3.ความเสี่ยงเชิงระบบต่อสังคม

International AI Safety Report 2026

1. ความเสี่ยงจากการใช้งานโดยเจตนาร้าย (Risks from Malicious Use)

1.1 เนื้อหาที่สร้างโดย AI กับอาชญากรรม (AI-generated Content & Criminal Activity)

- AI ถูกใช้สร้างข่าวปลอม (Fake news), การหลอกลวง (Scam), และการปลอมแปลงตัวตน (Impersonation)
- เพิ่มความเร็วและขนาดของอาชญากรรมดิจิทัล

1.2 การชักจูงและบิดเบือนความคิดเห็น (Influence & Manipulation)

- ใช้ AI เพื่อบิดเบือนความเห็นสาธารณะ
- ส่งผลต่อการเมือง การเลือกตั้ง และความเชื่อของสังคม

1.3 การโจมตีทางไซเบอร์ (Cyberattacks)

- AI ช่วยเพิ่มประสิทธิภาพการโจมตี เช่น Phishing, Malware, Social Engineering
- ลดต้นทุนและทักษะที่ผู้โจมตีต้องใช้

1.4 ความเสี่ยงด้านชีวภาพและเคมี (Biological & Chemical Risks)

- AI อาจถูกใช้ช่วยออกแบบหรือให้ข้อมูลอันตราย
- เพิ่มความเสี่ยงต่อความมั่นคงของมนุษยชาติ

International AI Safety Report 2026

2. ความเสี่ยงจากความผิดพลาดของระบบ (Risks from Malfunctions)

2.1 ความน่าเชื่อถือของระบบ (Reliability Challenges)

- AI อาจให้คำตอบผิดพลาดแต่ดูน่าเชื่อถือ (Confidently Wrong)
- ส่งผลร้ายเมื่อใช้ในบริบทสำคัญ เช่น การแพทย์ กฎหมาย ความมั่นคง

2.2 การสูญเสียการควบคุมของมนุษย์ (Loss of Control)

- ระบบ AI อาจตัดสินใจหรือกระทำเกินกว่าที่มนุษย์ตั้งใจ
- ความเสี่ยงเพิ่มขึ้นเมื่อ AI มีความอัตโนมัติสูง

3. ความเสี่ยงเชิงระบบต่อสังคม (Systemic Risks)

3.1 ผลกระทบต่อตลาดแรงงาน

(Labour Market Impacts)

งานบางประเภทถูกแทนที่อย่างรวดเร็ว

ช่องว่างทักษะ (Skill Gap) และความเหลื่อมล้ำเพิ่มขึ้น

3.2 ความเสี่ยงต่ออำนาจการตัดสินใจของมนุษย์

(Risks to Human Autonomy)

มนุษย์พึ่งพา AI มากเกินไป

การตัดสินใจสำคัญอาจถูกครอบงำโดยระบบอัตโนมัติ

Key Message

AI ไม่ได้เสี่ยงเพราะ “ฉลาดเกินไป”

แต่เสี่ยงเพราะ “ถูกใช้ผิด / ควบคุมไม่ดี / กระทบทั้งระบบสังคม”

ดังนั้น การกำกับดูแล AI ต้องครอบคลุม

- **ความมั่นคง (Security)**
- **ความน่าเชื่อถือ (Reliability)**
- **ธรรมาภิบาลและผลกระทบเชิงสังคม (Governance & Societal Impact)**

With inherent threats & Risks, AI might be Out of control if Lack to Security Governance

Security threats are evolving as AI technologies develop rapidly, and businesses are increasingly concerned about potential risks. Media sources suggest more companies are deploying open-source AI models in private environments. **However, nearly 90% of these deployments lack security mechanisms, posing significant risks**.

Risks with AI

Threats to computing environments

Attacks on AI computing environments threaten and undermine the foundation of the AI technology stack

- Vulnerability exploit
- Code attack
- Data breach
- Network attack
- Hardware attack
- Container attack
- Supply chain attack
- Other

Threats to AI systems

New attack threats to AI data, models, and prompts

- Data poisoning
- Adversarial examples
- Prompt injection
- Model extraction
- Model vulnerability
- Model inversion
- Membership inference
- Other

Threats resulting from the misuse of applications and services

Conscious or unconscious misuse of AI

Deepfake: AI is misused for fraud, misinformation, and identity forgery, causing large-scale data and privacy breaches.

Algorithm bias: AI automated decision-making causes bias against specific individuals or groups, and harms their interests.

Huawei's Mission for Cyber Security and Privacy Protection

"Huawei firmly believes that in the era of digitalization and intelligence, cyber security and privacy protection are the cornerstones of the development of the digital world. We persistently build cyber security and privacy protection capabilities into our products and services, and comply with applicable laws and regulations on cyber security and privacy protection.

Huawei guarantees that cyber security will never be outweighed by the consideration of commercial interests".

For more than 30 years, Huawei has served more than 3 billion people around the world and supported the stable operation of more than 1500 carrier networks in over 170 countries and regions. We have maintained a solid track record in cyber security, and our practices in cyber security have been widely recognized by customers.

Huawei's Cyber Security Journey

Network / Information-based

Cyber Security 1.0

2010–2017

Establish a compliance assurance and build a trust mechanism

- Establish Cyber Security Assurance System Incorporation of Product Security Redline/Baselines into Processes （SDL） Independent Verification System （ICSL)
- Delivery and Maintenance Operation Management
- Sales control
- Transparency Center

Digital

Cyber Security 2.0

2018–2024

Improving software engineering capabilities, build secure, trustworthy, and high-quality products

- 5G Cyber Security (NESAS/SCAS/MCKB)
- Software Engineering Capability Improvement (systematic enhancement of software quality, consolidation of software supply chain security management, etc.)
- Cyber Resilience & Native Security in Product
- Risk Mitigation & E2E Vulnerability Mgmt
- Hardware Security
- Cloud Operation Security
- Security Product & Solutions & Service

Intelligent

Cyber Security 3.0

2025–2030

Build cyber security and privacy competitiveness and continuously to be reliable partner in the intelligent era

- **5GA/6G Cyber Security**
- **Ecosystem Security** (HarmonyOS, Kunpeng, Ascend, and cloud…)
- **Computing Security** (Root of Trust, trusted boot, trusted computing and confidential computing)
- **AI Security** (Celia, Pangu, Hisec, ADN, etc.)
- **Data Security** (security engineering, etc.)
- **Cyber-physical System** (intelligent and digital applications such as smart city; safety)
- **Post-quantum Cryptography Algorithms** (technology, standards, engineering, etc.)

Outline

- บทที่ 1: บทนำ
- บทที่ 2: หลักการพื้นฐาน มาตรฐาน และภาพรวมภัยคุกคามด้านความมั่นคงปลอดภัย ระบบปัญญาประดิษฐ์
- บทที่ 3: กรอบการรักษาความมั่นคงปลอดภัยระบบปัญญาประดิษฐ์
- บทที่ 4: การกำกับดูแลและการบริหารความเสี่ยงสำหรับระบบปัญญาประดิษฐ์

Outline

ภาคผนวก

- ภาคผนวก ก. นิยามศัพท์ที่เกี่ยวข้องกับกระบวนการตรวจสอบและประเมินผล
- ภาคผนวก ข. ตัวอย่างรายการตรวจสอบความมั่นคงปลอดภัยสำหรับระบบปัญญาประดิษฐ์ (AI Security Checklist Example)
- ภาคผนวก ค. รายการตรวจสอบเพื่อการพัฒนาอย่างมั่นคงปลอดภัย (Secure Coding Checklist)
- ภาคผนวก ง: ตารางอ้างอิงภัยคุกคามและมาตรฐานควบคุมที่เกี่ยวข้อง (AI Threat and Control Framework Mapping)
- ภาคผนวก จ. เทมเพลตรายงานเหตุการณ์ฉุกเฉิน (AI-Specific Incident Response Template)
- ภาคผนวก ฉ. กรณีศึกษาการประยุกต์ใช้ปัญญาประดิษฐ์ในบริบทประเทศไทย
- ภาคผนวก ช. ตัวอย่าง นโยบายการใช้เทคโนโลยี Generative AI ที่ยอมรับได้ (Acceptable Use Policy: Generative AI)
- ภาคผนวก ซ. ตัวอย่าง นโยบายการใช้เทคโนโลยีระบบปัญญาประดิษฐ์ที่ยอมรับได้ (Acceptable Use Policy)

บทที่ 1: บทนำ

1.1 ความสำคัญของความมั่นคงปลอดภัย (Cybersecurity Importance)

- AI เพิ่มขีดความสามารถ/ประสิทธิภาพ แต่สร้างความเสี่ยงรูปแบบใหม่ (Novel Risks)
- ระบบ AI เป็นทั้งเป้าหมายและเครื่องมือของผู้โจมตี (Attack Target & Tool)
- ต้องยกระดับความน่าเชื่อถือ (Reliability) และความมั่นคงปลอดภัย (Security) ตลอดวงจรชีวิต (Lifecycle)

1.2 วัตถุประสงค์ของแนวปฏิบัติฯ (Objectives)

- แปลงหลักการเชิงนโยบายให้เป็นมาตรการปฏิบัติ (Actionable Controls) สำหรับองค์กรไทย
- ยกระดับความตระหนักรู้เรื่องภัยคุกคามไซเบอร์ที่จำเพาะต่อระบบ AI
- ส่งเสริมหลักการ Secure by Design, Secure by Default, Defense-in-Depth
- สอดคล้องกฎหมายไทย (Thai Regulations) เช่น พ.ร.บ.ไซเบอร์, พ.ร.บ.คุ้มครองข้อมูลส่วนบุคคล (PDPA)
- ยกระดับการรักษาความมั่นคงปลอดภัยไซเบอร์สำหรับ AI ของประเทศไทยให้ทัดเทียมกับระดับสากล

4.4 กรอบการทำงานเพื่อสนับสนุน ธรรมาภิบาลในการประยุกต์ใช้ AI

AI GOVERNANCE GUIDELINE

กรอบการทำงาน เพื่อสนับสนุนให้เกิด ธรรมาภิบาล ในการประยุกต์ใช้ AI ประกอบด้วย 3 องค์ประกอบหลัก ได้แก่

AI Governance Structure

- จัดตั้งคณะกรรมการกำกับดูแล (AI Governance Council) เพื่อกำหนดทิศทางการประยุกต์ใช้งาน AI ผ่านการกำหนดกลยุทธ์และนโยบาย รวมถึงเฝ้าติดตาม และประเมินผลกาประยุกต์ใช้ AI อย่างต่อเนื่อง เพื่อสนับสนุนให้เกิดธรรมาภิบาลในการประยุกต์ใช้ AI
- กำหนดหน้าที่ของบุคลากรและผู้มีส่วน-ได้เสียที่เกี่ยวข้องกับการประยุกต์ใช้ AI พร้อมทั้งสร้างความตระหนักรู้ในด้านความรับผิดชอบ (Responsibility) และความรับผิดชอบต่อผลของการกระทำ (Accountability) ของแต่ละหน้าที่
- พัฒนาศักยภาพของบุคลากรและผู้มีส่วนได้ส่วนเสียที่เกี่ยวข้อง เพื่อให้สามารถปฏิบัติงานได้อย่างเหมาะสมตามหน้าที่ที่ได้รับมอบหมาย

AI Strategy

- มองหาโอกาสในการนำ AI มาประยุกต์ใช้เพื่อสนับสนุนให้บรรลุเป้าหมายขององค์กรหรือเป้าหมายทางธุรกิจ
- กำหนดเป้าหมายในการประยุกต์ใช้ AI ตามลำดับความสำคัญ โดยพิจารณาจากประโยชน์ที่จะได้รับ ความพร้อมขององค์กร หลักการจริยธรรมปัญญาประดิษฐ์ กฎหมายและข้อกำหนดที่ต้องดำเนินการให้สอดคล้อง รวมถึง ความซับซ้อนและเวลาที่จำเป็นต้องใช้ในการดำเนินการ
- กำหนดกลยุทธ์ในการบริหารจัดการข้อมูล
- กำหนดแผนปฏิบัติงานในการประยุกต์ใช้ AI (AI Roadmap)
- วิเคราะห์ความเสี่ยงและผลกระทบที่อาจเกิดขึ้นจากการประยุกต์ใช้ AI รวมถึงการกำหนดระดับการมีส่วนร่วมของมนุษย์ในการทำงานของ AI และมาตรการในการควบคุมความเสี่ยงที่เหมาะสม เพื่อควบคุมความเสี่ยงให้อยู่ในขอบเขตที่ยอมรับได้

AI Operation

- จัดทำข้อกำหนดความต้องการในการพัฒนาระบบ AI (AI Requirement) พร้อมทั้งออกแบบโซลูชั่นที่เหมาะสมกับข้อกำหนดดังกล่าว
- จัดเตรียมข้อมูลที่มีคุณภาพสำหรับการสอน ตรวจสอบ และทดสอบโมเดลปัญญาประดิษฐ์ รวมถึงลดความมีอคติที่อาจเกิดจากข้อมูล (Data Bias)
- สร้างโมเดลปัญญาประดิษฐ์โดยนำหลักการจริยธรรมปัญญาประดิษฐ์มาปรับใช้ และควบคุมความเสี่ยงที่อาจเกิดขึ้น
- เฝ้าติดตามประสิทธิภาพ (Performance) การประยุกต์ใช้ AIรวมถึงการปฏิบัติงานตามนโยบาย หลักการจริยธรรมปัญญาประดิษฐ์ กฎหมายและข้อกำหนดที่เกี่ยวข้อง (Compliance)
- ประเมินผลการประยุกต์ใช้งาน AI ในปัจจุบันและกำหนดแนวทางการดำเนินงานอนาคต

1.4 ขอบเขต (Scope) และนอกขอบเขต (Out-of-Scope)

ในขอบเขต (In-Scope):

- การป้องกันการใช้ปัญญาประดิษฐ์ในทางที่ผิดเพื่อโจมตีระบบอื่น เช่น การสร้าง Deepfake เพื่อหลอกลวง หรือการใช้ปัญญาประดิษฐ์พัฒนามัลแวร์
- การปฏิบัติด้านความมั่นคงปลอดภัยตลอดวงจรชีวิตของระบบปัญญาประดิษฐ์ ตั้งแต่ การออกแบบ การพัฒนา การติดตั้งใช้งาน ไปจนถึงการดำเนินงานและบำรุงรักษา จนถึง การกำจัดและทำลาย โดยมุ่งเน้นที่การป้องกันระบบปัญญาประดิษฐ์จากการถูกโจมตี
- การป้องกันภัยคุกคามที่จำเพาะต่อระบบปัญญาประดิษฐ์ เช่น การโจมตีข้อมูล การโจมตีโมเดล และการโจมตีโครงสร้างพื้นฐาน
- การบริหารจัดการความเสี่ยงด้านความมั่นคงปลอดภัยไซเบอร์สำหรับปัญญาประดิษฐ์
- การรักษาความมั่นคงปลอดภัยของข้อมูล โมเดล แอปพลิเคชัน และโครงสร้างพื้นฐาน
- การประยุกต์ใช้กับระบบปัญญาประดิษฐ์ที่องค์กรพัฒนาขึ้นเอง ระบบที่นำส่วนประกอบ Open-Source และโมเดลที่ฝึกไว้ล่วงหน้า จากแหล่งภายนอกมาติดตั้ง ปรับจูน หรือใช้งาน

1.4 ขอบเขต (Scope) และนอกขอบเขต (Out-of-Scope)

นอกขอบเขต (Out-of-Scope):

- ประเด็นด้านจริยธรรม ความเท่าเทียม ความโปร่งใส และความเป็นธรรมของปัญญาประดิษฐ์ ซึ่งได้กล่าวถึงอย่างละเอียดแล้วในเอกสารที่จัดทำโดย ศูนย์ธรรมาภิบาลปัญญาประดิษฐ์
- ปัญญาประดิษฐ์เพื่อการป้องกันความมั่นคงปลอดภัย กล่าวคือการประยุกต์ใช้เทคโนโลยีปัญญาประดิษฐ์ในการระบุ ป้องกัน ตรวจจับ วิเคราะห์ และตอบสนองต่อภัยคุกคามในเชิงรุก ครอบคลุมสภาพแวดล้อมทางกายภาพ ดิจิทัล และแบบผสมผสาน โดยมีเป้าหมายเพื่อเพิ่มความเร็ว ความแม่นยำ ประสิทธิภาพ และความสามารถในการขยายขนาดของการปฏิบัติการด้านความมั่นคงปลอดภัยให้เหนือกว่าขีดความสามารถของมนุษย์

บทที่ 2:หลักการพื้นฐาน มาตรฐาน และภาพรวมภัยคุกคามด้านความมั่นคงปลอดภัย ระบบปัญญาประดิษฐ์

2.5 ความเสี่ยงจากการใช้ AI ที่ไม่มั่นคงปลอดภัย (Risks)

- อคติ (Bias), ความไม่โปร่งใส (Opacity), การละเมิดความเป็นส่วนตัว (Privacy)
- ความผิดพลาดในบริบทสำคัญ (Safety-Critical) เช่น การแพทย์/ยานยนต์อัตโนมัติ
- การรั่วไหลข้อมูลส่วนบุคคล (Data Leakage) และเหตุการณ์ไซเบอร์ (Cyber Incidents)

2.6 ภาพรวมภัยคุกคามด้านความมั่นคงปลอดภัยของปัญญาประดิษฐ์

ประเภทภัยคุกคาม	ผลกระทบต่อ C-I-A	เหตุผล
๑. Data Poisoning	Integrity	เป็นการบ่อนทำลายความถูกต้องของโมเดลตั้งแต่กระบวนการฝึกสอน ทำให้ผลลัพธ์ขาดความน่าเชื่อถือ
๒. Data Privacy Breaches	Confidentiality	มีเป้าหมายเพื่อเปิดเผยข้อมูลที่ละเอียดอ่อนจากชุดข้อมูลฝึกสอน ซึ่งเป็นการละเมิดความลับของข้อมูลโดยตรง
๓. Evasion	Integrity	ทำให้โมเดลจำแนกประเภทผิดพลาด ณ ขณะใช้งาน ซึ่งเป็นการลดทอนความถูกต้องและความน่าเชื่อถือของผลลัพธ์
๔. Prompt Injection	Integrity, Confidentiality	การสั่งงานนอกขอบเขตกระทบต่อความถูกต้อง (I) และอาจนำไปสู่การเปิดเผยข้อมูลที่เป็นความลับ (C)
๕. Model Extraction	Confidentiality	เป็นการลอกเลียนแบบเพื่อขโมยทรัพย์สินทางปัญญา ซึ่งถือเป็นข้อมูลที่เป็นความลับขององค์กร
๖. Model Poisoning	Integrity	มีเป้าหมายเพื่อบิดเบือนการทำงานและผลลัพธ์ของโมเดลโดยตรง ทำให้ขาดความถูกต้องสมบูรณ์
๗. Denial of Service (DoS)	Availability	มุ่งเป้าให้ระบบหยุดทำงานหรือทำงานช้าลง ทำให้ผู้ใช้ที่ได้รับอนุญาตไม่สามารถเข้าถึงบริการได้
๘. AI Supply Chain	Confidentiality, Integrity, Availability	ส่วนประกอบที่มีช่องโหว่สามารถนำไปสู่การขโมยข้อมูล (C) การบิดเบือนการทำงาน (I) หรือทำให้ระบบล่ม (A)
๙. Infrastructure Attacks	Confidentiality, Integrity, Availability	ช่องโหว่ในระบบแวดล้อมอาจนำไปสู่การเข้าถึงข้อมูล โดยไม่ได้รับอนุญาต (C) การแก้ไขผลลัพธ์ (I) และ การทำให้ระบบหยุดชะงัก (A)
๑๐. Direct Model Theft	Confidentiality	เป็นการขโมยไฟล์โมเดลโดยตรงซึ่งเป็นการละเมิดทรัพย์สินทางปัญญาที่เป็นความลับขององค์กร

2.7 มาตรฐานและกรอบ (Standards & Frameworks)

- ISO/IEC 42001 (AI Management System), ISO/IEC 23894 (AI Risk Management)
- ISO/IEC 27001/27002 (ISMS), ISO/IEC 5338 (AI System Life Cycle Processes)
- ENISA Multilayer Framework (Cybersecurity Foundations / AI-Specific / Sectoral)
- OWASP Top 10 for LLM Applications, OWASP AI Exchange

2.8 กฎหมายไทยที่เกี่ยวข้อง (Thai Regulations)

- พ.ร.บ. การรักษาความมั่นคงปลอดภัยไซเบอร์ พ.ศ. 2562 (Cybersecurity Act) — CII/มาตรการควบคุม
- พ.ร.บ. คุ้มครองข้อมูลส่วนบุคคล พ.ศ. 2562 (PDPA) — ฐานกฎหมาย (Legal Basis), สิทธิของเจ้าของข้อมูล (Data Subject Rights), มาตรการรักษาความปลอดภัย (Security Measures)
- พ.ร.บ. ว่าด้วยการกระทำความผิดเกี่ยวกับคอมพิวเตอร์ พ.ศ. ๒๕๕๐ และฉบับแก้ไขเพิ่มเติม เป็นกฎหมายหลักที่กำหนดฐานความผิดและบทลงโทษเกี่ยวกับการใช้ระบบคอมพิวเตอร์ในทางที่มิชอบ โดยมุ่งเน้นการป้องกันและปราบปรามการนำเข้าหรือเผยแพร่เนื้อหาที่ผิดกฎหมาย อันอาจส่งผลกระทบต่อความมั่นคงของประเทศและศีลธรรมอันดีของประชาชน
- ข้อพิจารณาพิเศษ การรักษาอธิปไตยทางข้อมูลในระบบปัญญาประดิษฐ์ นอกเหนือจากมาตรการความมั่นคงปลอดภัยเชิงเทคนิคแล้ว การกำกับดูแลระบบปัญญาประดิษฐ์ จำเป็นต้องคำนึงถึงมิติของ "อธิปไตยทางข้อมูล" อย่างเคร่งครัด
- แนวทางธรรมาภิบาลปัญญาประดิษฐ์ของ ETDA (AI Governance Guideline for Executives/Generative AI)

ข้อสรุปสำคัญ (Key Takeaways)

- เชื่อมโยงธรรมาภิบาล (Governance) กับความมั่นคงปลอดภัย (Security) ให้เป็นกระบวนการที่ปฏิบัติได้ (Actionable)
- ใช้แนวคิด Secure by Design/Default + Defense-in-Depth ครอบคลุมทั้ง Data/Model/App/Infra
- ใช้มาตรฐานสากล (ISO/NIST/OWASP/ENISA/CISA-NCSC) และปฏิบัติตามกฎหมายไทย (PDPA/พ.ร.บ.ไซเบอร์)
- เตรียมเครื่องมือทดสอบ/เฝ้าระวังจาก OWASP AI Exchange และประเมินความเสี่ยงตาม ISO/IEC 23894

บทที่ 3: กรอบการรักษาความมั่นคงปลอดภัยระบบปัญญาประดิษฐ์ (Secure AI Usage Guidelines)

Agenda

- หลักการพื้นฐาน — Secure by Design, Secure by Default, Defense-in-Depth
- กรอบวงจรชีวิตระบบ AI ที่บูรณาการความมั่นคงปลอดภัย (Security-Integrated AI Lifecycle)
- การเชื่อมโยงกับมาตรฐาน (ISO/IEC 22989, 27002, 42001, 23894, OWASP AI)
- สรุปกิจกรรมสำคัญและผลลัพธ์ที่คาดหวังในแต่ละระยะ

หลักการพื้นฐาน (Fundamental Principles)

- ออกแบบอย่างมั่นคงปลอดภัย (Secure by Design) — ความปลอดภัยเป็นข้อกำหนดหลักตั้งแต่ต้น (SDLC)
- ตั้งค่าเริ่มต้นอย่างมั่นคงปลอดภัย (Secure by Default) — กำหนดค่าตั้งต้นให้มั่นคงปลอดภัยที่สุด
- การป้องกันเชิงลึก (Defense-in-Depth) — ควบคุมหลายชั้น (Layers) เพื่อกันพลาดชั้นเดียว

Secure by Design

- แบบจำลองภัยคุกคามและประเมินความเสี่ยง (Threat Modeling & Risk Assessment)
- การรวบรวม/จัดเตรียมข้อมูลอย่างปลอดภัย (Data Acquisition & Preparation) — ใช้เทคโนโลยียกระดับความเป็นส่วนตัว (Privacy-Enhancing Technologies: PETs) เช่น การนิรนาม (Anonymization), นามแฝง (Pseudonymization), ตรวจสอบแหล่งที่มา (Data Provenance)
- พัฒนา/ฝึกสอนโมเดลให้ทนทาน (Robustness) — ฝึกสอนเชิงปรปักษ์ (Adversarial Training)

Secure by Default

- ควบคุมการเข้าถึง (Access Control) ตามหลักสิทธิ์น้อยที่สุด (Principle of Least Privilege: PoLP)
- พิสูจน์ตัวตนหลายปัจจัย (Multi-Factor Authentication: MFA) สำหรับผู้ดูแลระบบ (Admin)
- ทำความสะอาดข้อมูลนำเข้า (Input Sanitization) — ป้องกัน Prompt Injection / Data Manipulation
- ตั้งค่า API ให้เข้มงวด (AuthN/AuthZ, Rate Limiting) — กัน Model Extraction
- บันทึกเหตุการณ์และเฝ้าระวัง (Logging & Monitoring) — ตรวจ Data Exfiltration, Adversarial Attack

Defense-in-Depth — การประยุกต์ใช้หลายชั้น (Layers)

- เครือข่าย (Network Layer) — WAF, แบ่งส่วนเครือข่าย (Network Segmentation)
- แอปพลิเคชัน (Application Layer) — AuthN/AuthZ เข้มงวด, Input Sanitization
- โมเดล (Model Layer) — Adversarial Training, Output Filtering/Guardrails
- ข้อมูล (Data Layer) — เข้ารหัส (Encryption) ทั้ง at-rest / in-transit, ควบคุมการเข้าถึงชุดข้อมูล
- การปฏิบัติการและการเฝ้าระวัง(Operational & Monitoring Layer) — Logging, Anomaly Detection

Huawei is committed to promoting the collaborative progress of AI development and governance

We believe that

AI should **promote the human, societal, and environmental well-being, protect privacy, and remain safe, controllable, transparent, explainable, collaborative, and sustainable**.

Huawei **values AI technology innovation and governance responsibilities. We are dedicated to responsibly developing and deploying AI products in compliance with laws, promoting inclusive AI, and working with partners to address the challenges encountered during societal development**.

Governance and development are interdependent.

中国人工智能产业发展联盟 Artificial Intelligence Industry Alliance

AI Subcommittee of the NITS Technical Committee

中国人工智能学会 Chinese Association for Artificial Intelligence

AI Assurance Club — A Global Digital Foundation Initiative

ISO/IEC JTC 1/SC 42 Artificial intelligence

WORLD ECONOMIC FORUM

HUAWEI

กรอบวงจรชีวิต AI ที่บูรณาการความมั่นคงปลอดภัย (Security-Integrated Lifecycle)

- ระยะที่ 0: แนวคิด (Concept)
- ระยะที่ 1: ออกแบบอย่างมั่นคงปลอดภัย (Secure Design)
- ระยะที่ 2: พัฒนาอย่างมั่นคงปลอดภัย (Secure Development)
- ระยะที่ 3: ทวนสอบด้านความมั่นคงปลอดภัย (Secure Verification)
- ระยะที่ 4: นำไปใช้งานอย่างมั่นคงปลอดภัย (Secure Deployment)
- ระยะที่ 5: ดำเนินงานและบำรุงรักษาอย่างมั่นคงปลอดภัย (Secure Operation & Maintenance)
- ระยะที่ 6: กำจัดและทำลายอย่างมั่นคงปลอดภัย (Disposal)

การเชื่อมโยงกับ ISO/IEC 22989 (Functional ⟶Security-Integrated)

ISO/IEC 22989:2022 (Functional View)	กรอบวงจรชีวิตของระบบปัญญาประดิษฐ์ ที่นำเสนอ (Security-Integrated View)	คำอธิบายการเชื่อมโยง
1. Inception	ระยะที่ 0: ขั้นตอนแนวคิด (Concept)	เป็นขั้นตอนเดียวกัน คือการระบุความต้องการและเป้าหมาย
2. Design and development	ระยะที่ 1-2: การออกแบบอย่างมั่นคงปลอดภัยและการพัฒนาอย่างมั่นคงปลอดภัย (Secure Design & Secure Development)	แยกขั้นตอนเดียวของ ISO เพื่อเน้นย้ำกิจกรรมความปลอดภัยเชิงรุก (Shift-Left)
3. Verification and validation	ระยะที่ 3: การทวนสอบด้านความมั่นคงปลอดภัย (Secure Verification)	เป็นขั้นตอนเดียวกัน โดยเน้นการทดสอบด้านความมั่นคงปลอดภัยโดยเฉพาะ
4. Deployment	ระยะที่ 4: การนำไปใช้งานอย่างมั่นคงปลอดภัย (Secure Deployment)	เป็นขั้นตอนเดียวกัน โดยเน้นการตั้งค่าที่ปลอดภัยและการส่งมอบที่รัดกุม
5. Operation and monitoring 6. Re-evaluation	ระยะที่ 5: การดำเนินงานและการบำรุงรักษาอย่างมั่นคงปลอดภัย (Secure Operation & Maintenance)	รวม 2 ขั้นตอนของ ISO เข้าด้วยกัน เนื่องจากในทางปฏิบัติ การเฝ้าระวังและประเมินซ้ำเป็นส่วนหนึ่งของการดำเนินงานและบำรุงรักษาอย่างต่อเนื่อง
7. Retirement	ระยะที่ 6: กระบวนการกำจัดและทำลาย (Disposal)	เป็นขั้นตอนเดียวกัน โดยใช้คำว่า "Disposal" เพื่อเน้นย้ำถึงการทำลายข้อมูลอย่างปลอดภัย

ระยะที่ 0: แนวคิด (Concept)

- การพิจารณาบริบททางกฎหมายและข้อบังคับ
- กำหนดสินทรัพย์-ภัยคุกคาม-ช่องโหว่ (Asset-Threat-Vulnerability)
- ประเมินความเสี่ยง (Risk Assessment: ISO/IEC 23894) และภูมิทัศน์ภัยคุกคาม (Threat Landscape)
- ผลลัพธ์: ทะเบียนความเสี่ยง (Risk Register), ข้อกำหนดความปลอดภัยเบื้องต้น (Initial Security Requirements), AI System Impact Assessment (ISO/IEC 42001)

ระยะที่ 0: แนวคิด (Concept)

1.การพิจารณาบริบททางกฎหมายและข้อบังคับ

1.1) กฎหมายและกฎระเบียบภายในประเทศไทย

1.1.1) พระราชบัญญัติการรักษาความมั่นคงปลอดภัยไซเบอร์ พ.ศ. 2562 หากระบบปัญญาประดิษฐ์ถูกนำไปใช้ในหน่วยงานที่เป็นโครงสร้างพื้นฐานสำคัญทางสารสนเทศ องค์กรจะต้องปฏิบัติตามมาตรฐานขั้นต่ำและกระบวนการที่กำหนดอย่างเข้มงวด

1.1.2) กฎระเบียบเฉพาะทาง พิจารณากฎระเบียบของหน่วยงานกำกับดูแลใน

แต่ละอุตสาหกรรม เช่น ธนาคารแห่งประเทศไทยสำหรับ FinTech หรือสำนักงานคณะกรรมการกำกับหลักทรัพย์และตลาดหลักทรัพย์ (ก.ล.ต.) สำหรับการใช้ปัญญาประดิษฐ์ในการซื้อขายสินทรัพย์ดิจิทัล

1.2) ข้อกำหนดของคู่สัญญาและกฎหมายระหว่างประเทศ

1.2.1) ข้อตกลงในสัญญา สัญญาทางธุรกิจกับคู่ค้าอาจมีข้อกำหนดเฉพาะเกี่ยวกับระดับความมั่นคงปลอดภัย การจัดการข้อมูล หรือมาตรฐานด้านจริยธรรมของปัญญาประดิษฐ์ที่องค์กรต้องปฏิบัติตาม

1.2.2) General Data Protection Regulation (GDPR) หากระบบปัญญาประดิษฐ์มีการประมวลผลข้อมูลส่วนบุคคลของพลเมืองในสหภาพยุโรป หรือมีคู่สัญญาที่ดำเนินธุรกิจในสหภาพยุโรป องค์กรจำเป็นต้องปฏิบัติตามข้อกำหนดของ GDPR ซึ่งมีความเข้มงวดสูง

1.2.3) EU AI Act แม้จะเป็นกฎหมายของ EU แต่มีผลกระทบข้ามพรมแดนองค์กรควรพิจารณาแนวทางการจำแนกความเสี่ยง ดังแสดงในหัวข้อ 2.3 เพื่อเตรียมความพร้อมและสร้างความได้เปรียบในการแข่งขัน

ระยะที่ 0: แนวคิด (Concept)

2.การทำความเข้าใจองค์ประกอบของความเสี่ยง: Asset, Threat, Vulnerability

- สินทรัพย์ (Asset) คือสิ่งใดก็ตามที่มีคุณค่าต่อองค์กรซึ่งเกี่ยวข้องกับระบบปัญญาประดิษฐ์ สินทรัพย์เหล่านี้อาจเป็นได้ทั้งสิ่งที่จับต้องได้และจับต้องไม่ได้ ตัวอย่างเช่น
  - โมเดลปัญญาประดิษฐ์ที่ฝึกสอนแล้ว (Trained AI Model) ถือเป็นทรัพย์สินทางปัญญา (Intellectual Property) ที่มีมูลค่าสูง อาจเป็นเป้าหมายของการโจมตีเพื่อขโมย (Model Extraction)
  - ชุดข้อมูลฝึกสอน (Training Dataset) โดยเฉพาะอย่างยิ่งหากมีข้อมูลที่ละเอียดอ่อน เช่น ข้อมูลส่วนบุคคล (PII) ข้อมูลทางการเงิน หรือความลับทางการค้า การรั่วไหลของข้อมูลนี้อาจนำไปสู่การละเมิดความเป็นส่วนตัวอย่างรุนแรง
  - ชื่อเสียงและความน่าเชื่อถือขององค์กร (Organizational Reputation) การที่ระบบปัญญาประดิษฐ์ตัดสินใจผิดพลาดอย่างร้ายแรง มีอคติ หรือสร้างเนื้อหาที่เป็นอันตราย อาจทำลายความไว้วางใจของลูกค้าและสาธารณชน

ระยะที่ 0: แนวคิด (Concept)

2.การทำความเข้าใจองค์ประกอบของความเสี่ยง: Asset, Threat, Vulnerability

- ภัยคุกคาม (Threat) คือเหตุการณ์หรือผู้กระทำ (Actor) ที่มีศักยภาพในการสร้างความเสียหายต่อสินทรัพย์
  - ➢ผู้ไม่หวังดีภายนอก (External Malicious Actor) แฮกเกอร์ที่พยายามโจมตีระบบเพื่อผลประโยชน์ทางการเงินหรือเพื่อสร้างชื่อเสียง
  - ➢การโจมตีเชิงปรปักษ์ (Adversarial Attack) เป็นภัยคุกคามในรูปแบบของเทคนิคที่ถูกสร้างขึ้นเพื่อหลอกลวงหรือบิดเบือนการทำงานของโมเดลปัญญาประดิษฐ์โดยเฉพาะ
  - ➢ผู้ใช้งานภายในที่ขาดความตระหนัก (Unaware Insider) พนักงานที่อาจสร้างพรอมต์ (Prompt) ที่นำไปสู่การเปิดเผยข้อมูลที่ละเอียดอ่อนโดยไม่ได้ตั้งใจ

ระยะที่ 0: แนวคิด (Concept)

2.การทำความเข้าใจองค์ประกอบของความเสี่ยง: Asset, Threat, Vulnerability

- ช่องโหว่ (Vulnerability) คือจุดอ่อนหรือข้อบกพร่องในระบบ กระบวนการ หรือมาตรการควบคุม ที่เปิดโอกาสให้ภัยคุกคามสามารถสร้างความเสียหายต่อสินทรัพย์ได้สำเร็จ ตัวอย่างในบริบทปัญญาประดิษฐ์ เช่น
  - การขาดการฝึกสอนโมเดลให้ทนทาน (Lack of Adversarial Training) ทำให้โมเดลไม่มีภูมิต้านทานและอ่อนไหวต่อการโจมตีแบบ Evasion Attacks
  - การขาดการตรวจสอบและคัดกรองข้อมูลนำเข้า (Insufficient Input Sanitization) เปิดช่องให้เกิดการโจมตีแบบ Prompt Injection ในระบบ LLM
  - API ที่ไม่มีการป้องกัน การเปิดเผย API ของโมเดลโดยไม่มีการพิสูจน์ตัวตนหรือการจำกัดอัตราการเรียกใช้ (Rate Limiting) ทำให้ง่ายต่อการโจมตีแบบ Model Extraction

ระยะที่ 0: แนวคิด (Concept)

3. กระบวนการประเมินความเสี่ยงสำหรับระบบปัญญาประดิษฐ์ (AI System Risk Assessment)

3.1 การระบุความเสี่ยง (Risk Identification)

- เป็นกระบวนการเชิงรุกในการจำแนกและจัดทำรายการภัยคุกคามและช่องโหว่ที่อาจส่งผลกระทบต่อสินทรัพย์ของระบบปัญญาประดิษฐ์ โดยพิจารณาจากแหล่งข้อมูลต่างๆ เช่น ISO/IEC 23894:2023 Annex B และ OWASP AI Security Top 10
- ตัวอย่าง: ระบบปัญญาประดิษฐ์ประเมินสินเชื่อ
  - สินทรัพย์: ความถูกต้องของการอนุมัติสินเชื่อ ข้อมูลทางการเงินของผู้สมัคร ชื่อเสียงของสถาบันการเงิน
  - การระบุความเสี่ยง: ผู้ไม่หวังดีอาจใช้เทคนิค Data Poisoning โดยการแทรกข้อมูลผู้สมัครปลอมที่มีลักษณะดีแต่แฝงความสัมพันธ์ที่ผิดปกติเข้าไปในชุดข้อมูลฝึกสอน (อ้างอิง OWASP-AI-EXCHANGE 3. Development-time threats) เพื่อสร้าง Backdoor ทำให้โมเดลมีแนวโน้มที่จะอนุมัติสินเชื่อให้กับผู้สมัครที่มีลักษณะคล้ายกันในอนาคต แม้ว่าจะมีคุณสมบัติไม่เพียงพอก็ตาม

ระยะที่ 0: แนวคิด (Concept)

3. กระบวนการประเมินความเสี่ยงสำหรับระบบปัญญาประดิษฐ์ (AI System Risk Assessment)

3.2 การวิเคราะห์ความเสี่ยง (Risk Analysis)

เป็นขั้นตอนการประเมินโอกาสการเกิด (Likelihood) และผลกระทบ (Impact) ของความเสี่ยงที่ระบุไว้ เพื่อทำความเข้าใจระดับความรุนแรงของแต่ละความเสี่ยง โดยมีตัวอย่างดังนี้

- การวิเคราะห์โอกาสเกิด (Likelihood) อาจอยู่ในระดับ “ปานกลาง” (Medium) เนื่องจากการโจมตีประเภทนี้ต้องอาศัยการเข้าถึงหรือมีอิทธิพลต่อกระบวนการรวบรวมข้อมูลซึ่งอาจทำได้ยาก แต่หากทำสำเร็จจะตรวจจับได้ยากมาก
- การวิเคราะห์ผลกระทบ (Impact) อยู่ในระดับ “สูง” (High) เนื่องจากหากการโจมตีสำเร็จจะนำไปสู่ความเสียหายทางการเงินโดยตรง (หนี้เสีย) การละเมิดกฎเกณฑ์ของหน่วยงานกำกับดูแล และการสูญเสียความน่าเชื่อถือขององค์กร (อ้างอิง ISO/IEC 42001:2023 A.5 Assessing impacts of AI systems)

ระยะที่ 0: แนวคิด (Concept)

3. กระบวนการประเมินความเสี่ยงสำหรับระบบปัญญาประดิษฐ์ (AI System Risk Assessment)

3.3 การประเมินระดับความเสี่ยง (Risk Evaluation)

- เป็นขั้นตอนสุดท้ายในการเปรียบเทียบผลลัพธ์จากการวิเคราะห์ความเสี่ยงกับเกณฑ์ความเสี่ยงที่องค์กรยอมรับได้ (Risk Acceptance Criteria) เพื่อจัดลำดับความสำคัญและตัดสินใจว่าความเสี่ยงใดที่ต้องมีมาตรการจัดการ
- โดยมีตัวอย่างเช่น การประเมินระดับความเสี่ยง จากผลการวิเคราะห์ (โอกาสเกิด “ปานกลาง” x ผลกระทบ “สูง”) ทำให้ความเสี่ยงโดยรวมอยู่ในระดับ “สูง” (High) ซึ่งเกินกว่าระดับความเสี่ยงที่องค์กรยอมรับได้
- ดังนั้น ความเสี่ยงนี้จึงต้องได้รับการจัดการ โดยต้องกำหนดมาตรการควบคุมในระยะที่ 1 (Secure Design) เช่น การออกแบบกระบวนการตรวจสอบความถูกต้องและแหล่งที่มาของข้อมูล (Data Provenance) อย่างเข้มงวด

ระยะที่ 0: แนวคิด (Concept)

4. ผลลัพธ์ที่คาดหวัง (Expected Outcomes)

เมื่อสิ้นสุดระยะที่ 0 องค์กรควรมีเอกสารและข้อกำหนดที่ชัดเจน ซึ่งเป็นผลลัพธ์จากการประเมินความเสี่ยง ได้แก่

4.1) สรุปข้อกำหนดทางกฎหมายและข้อบังคับที่เกี่ยวข้อง

4.2) ทะเบียนความเสี่ยง เอกสารที่รวบรวมความเสี่ยงทั้งหมดของระบบปัญญาประดิษฐ์ที่ระบุและประเมินไว้

4.3) ข้อกำหนดด้านความมั่นคงปลอดภัยเบื้องต้น ข้อกำหนดที่ต้องถูกนำไปพิจารณาในขั้นตอนการออกแบบ เช่น "ระบบต้องมีกลไกป้องกันการโจมตีแบบ Prompt Injection" หรือ "ข้อมูลส่วนบุคคลทั้งหมดในชุดข้อมูลต้องผ่านกระบวนการทำข้อมูลนิรนาม"

4.4) เอกสารประเมินผลกระทบของระบบปัญญาประดิษฐ์ตามแนวทางของ ISO/IEC 42001:2023 เพื่อทำความเข้าใจผลกระทบในวงกว้างต่อบุคคลและสังคม

ระยะที่ 1: ออกแบบอย่างมั่นคงปลอดภัย (Secure Design)

- ทำ Threat Modeling สำหรับส่วนประกอบ AI (STRIDE ประยุกต์กับ Data/Model/API)
- ออกแบบสถาปัตยกรรมที่มั่นคงปลอดภัย (Secure Architecture) — Isolation, Detection, Failsafe, Redundancy
- ผลลัพธ์: Threat Model Document, Secure Architecture Diagram/Specs, Detailed Security Requirements

ระยะที่ 1: ออกแบบอย่างมั่นคงปลอดภัย (Secure Design)

1. การสร้างแบบจำลองภัยคุกคามสำหรับ AI (Threat Modeling for AI)

- การสร้างแบบจำลองภัยคุกคามเป็นกระบวนการเชิงรุกที่เป็นระบบ เพื่อวิเคราะห์ ระบุ และจัดลำดับความสำคัญของภัยคุกคามที่อาจเกิดขึ้นกับระบบปัญญาประดิษฐ์ การทำ Threat Modeling ในบริบทของปัญญาประดิษฐ์ จะต้องพิจารณาพื้นผิวการโจมตี (Attack Surface) ที่เป็นเอกลักษณ์ของเฉพาะเจาะจงต่อวงจรชีวิตของปัญญาประดิษฐ์ (AI Lifecycle) ทั้งหมด ตั้งแต่ข้อมูลไปจนถึงโมเดลและ API
- กระบวนการนี้สามารถใช้กรอบการวิเคราะห์ภัยคุกคาม เช่น STRIDE (Spoofing, Tampering, Repudiation, Information Disclosure, Denial of Service, Elevation of Privilege) มาประยุกต์ใช้กับส่วนประกอบต่างๆ ของปัญญาประดิษฐ์ได้ดังนี้

ระยะที่ 1: ออกแบบอย่างมั่นคงปลอดภัย (Secure Design)

1. การสร้างแบบจำลองภัยคุกคามสำหรับ AI (Threat Modeling for AI)

- ตัวอย่างการประยุกต์ใช้ Threat Modeling กับระบบปัญญาประดิษฐ์ แนะนำผลิตภัณฑ์
  - ส่วนประกอบ ไปป์ไลน์ข้อมูล (Data Pipeline)
    - ภัยคุกคาม (Tampering) ผู้ไม่หวังดีหรือคู่แข่งอาจแทรกข้อมูลการซื้อหรือรีวิวปลอม เข้ามาในชุดข้อมูลฝึกสอน (Data Poisoning) เพื่อบิดเบือนให้โมเดลแนะนำผลิตภัณฑ์ของตนเองมากกว่าของคู่แข่ง
    - มาตรการควบคุมเชิงออกแบบ (Design Control) ออกแบบระบบ Data Provenance เพื่อตรวจสอบแหล่งที่มาและความสมบูรณ์ของข้อมูล กำหนดให้มีขั้นตอน Data Validation and Sanitization ที่เข้มงวดก่อนนำข้อมูลเข้าสู่กระบวนการฝึกสอน (อ้างอิง OWASP-AI-EXCHANGE #DEVDATAPROTECT)
  - ส่วนประกอบ โมเดลปัญญาประดิษฐ์ (AI Model)
    - ภัยคุกคาม (Information Disclosure) โมเดลอาจจดจำและเปิดเผยข้อมูลส่วนบุคคล (PII) ของลูกค้าที่อยู่ในชุดข้อมูลฝึกสอนโดยไม่ได้ตั้งใจ ผ่านผลลัพธ์การแนะนำผลิตภัณฑ์
    - มาตรการควบคุมเชิงออกแบบ กำหนดในสถาปัตยกรรมให้ต้องใช้เทคนิค Privacy-Preserving Machine Learning เช่น Differential Privacy ในระหว่างกระบวนการฝึกสอน เพื่อลดความเสี่ยงที่โมเดลจะจดจำข้อมูลที่ละเอียดอ่อน

ระยะที่ 1: ออกแบบอย่างมั่นคงปลอดภัย (Secure Design)

1. การสร้างแบบจำลองภัยคุกคามสำหรับ AI (Threat Modeling for AI)

- ตัวอย่างการประยุกต์ใช้ Threat Modeling กับระบบปัญญาประดิษฐ์ แนะนำผลิตภัณฑ์
  - ส่วนประกอบ API สำหรับให้บริการ (Serving API)
    - ภัยคุกคาม (Denial of Service / Information Disclosure) ผู้โจมตีส่งคำร้อง (Query) ไปยัง API จำนวนมหาศาลเพื่อพยายามทำวิศวกรรมย้อนกลับและขโมยโมเดล (Model Extraction) หรือทำให้ระบบล่ม
    - มาตรการควบคุมเชิงออกแบบ ออกแบบ API ให้บังคับใช้กลไก การยืนยันตัวตน (Authentication) การอนุญาตสิทธิ์ (Authorization) และที่สำคัญคือ การจำกัดอัตราการเรียกใช้งาน (Rate Limiting) อย่างเข้มงวด

ระยะที่ 1: ออกแบบอย่างมั่นคงปลอดภัย (Secure Design)

2. การออกแบบสถาปัตยกรรมที่มั่นคงปลอดภัย (Secure Architecture Design)

จากผลลัพธ์ของ Threat Model ขั้นตอนถัดมาคือการออกแบบสถาปัตยกรรมของระบบปัญญาประดิษฐ์ ที่มีความยืดหยุ่นและทนทาน (Resilient) ต่อการโจมตี โดยอาศัยหลักการทางวิศวกรรมและสถาปัตยกรรมที่มั่นคงปลอดภัย (อ้างอิง ISO/IEC 27002:2022 A8.27) ซึ่งมีหลักการสำคัญดังนี้

2.1) การออกแบบกลไกป้องกันผลลัพธ์ที่เป็นอันตราย

- สำหรับระบบ Generative AI ต้องมีการออกแบบกลไกควบคุม เช่น การกรองผลลัพธ์ เพื่อป้องกันไม่ให้ระบบถูกใช้เป็นเครื่องมือในการสร้างเนื้อหาที่เป็นอันตรายตามที่ผู้ใช้ร้องขอ เช่น โคดของมัลแวร์ หรือข้อความที่ใช้ในการหลอกลวง

2.2) การแยกส่วน

- 2.2.1) หลักการการแยกส่วนประกอบต่าง ๆ ของระบบออกจากกัน เพื่อจำกัดขอบเขตของความเสียหายหากส่วนใดส่วนหนึ่งถูกโจมตี
- 2.2.2) ตัวอย่างในสถาปัตยกรรมปัญญาประดิษฐ์ ออกแบบให้สภาวะแวดล้อมสำหรับการฝึกสอนโมเดล ซึ่งมีข้อมูลดิบที่ละเอียดอ่อน แยกขาดจากสภาวะแวดล้อมสำหรับการให้บริการโมเดลที่ต้องเชื่อมต่อกับภายนอก การบุกรุกที่ Inference API จะต้องไม่สามารถเข้าถึงสภาวะแวดล้อมสำหรับการฝึกสอนโมเดลได้

ระยะที่ 1: ออกแบบอย่างมั่นคงปลอดภัย (Secure Design)

2.3) การตรวจจับ

- 2.3.1) หลักการการมีกลไกสำหรับตรวจจับกิจกรรมที่น่าสงสัยหรือเป็นอันตราย
- 2.3.2) ตัวอย่างในสถาปัตยกรรมปัญญาประดิษฐ์ ออกแบบระบบเฝ้าระวังที่ไม่ได้ดูแค่การใช้ CPU และ Memory แต่สามารถตรวจจับความเบี่ยงเบนของข้อมูล หรือพฤติกรรมของโมเดลที่ผิดปกติ ซึ่งอาจเป็นสัญญาณของการโจมตีแบบ Data Poisoning ที่ค่อย ๆ เกิดขึ้น นอกจากนี้ยังต้องออกแบบให้สามารถตรวจจับรูปแบบการส่งคำร้องที่ผิดปกติซึ่งอาจบ่งชี้ถึงความพยายามทำ Model Extraction

2.4) กลไกป้องกันความเสียหายเมื่อระบบล้มเหลว

- 2.4.1) หลักการออกแบบให้ระบบเข้าสู่สถานะที่ปลอดภัยที่สุดเมื่อเกิดความล้มเหลวหรือตรวจพบการโจมตี
- 2.4.2) ตัวอย่างในสถาปัตยกรรมปัญญาประดิษฐ์สำหรับ LLM ที่ใช้ภายในองค์กร หากระบบตรวจจับได้ว่ามี Prompt Injection ที่พยายามเข้าถึงข้อมูลลับ กลไกป้องกันความเสียหายเมื่อระบบล้มเหลวที่ออกแบบไว้คือการตัดการทำงานและส่งคืนคำตอบมาตรฐานที่ปลอดภัย เช่น "ขออภัยไม่สามารถดำเนินการตามคำขอนี้ได้" แทนที่จะพยายามประมวลผลคำสั่งที่เป็นอันตรายนั้นต่อไป

ระยะที่ 1: ออกแบบอย่างมั่นคงปลอดภัย (Secure Design)

2.5) ระบบสำรอง

- 2.5.1) หลักการการมีส่วนประกอบสำรองเพื่อรับประกันความพร้อมใช้งานและความถูกต้องสมบูรณ์
- 2.5.2) ตัวอย่างในสถาปัตยกรรมปัญญาประดิษฐ์สำหรับระบบปัญญาประดิษฐ์ วินิจฉัยโรคที่มีความสำคัญสูง การออกแบบอาจกำหนดให้ใช้โมเดลแบบรวมกลุ่ม (Ensemble Models) ที่ประกอบด้วยโมเดลหลาย ๆ ตัวที่มีสถาปัตยกรรมหรือถูกฝึกสอนบนชุดข้อมูลที่แตกต่างกันเล็กน้อย หากโมเดลตัวหนึ่งถูกหลอกลวงโดยการโจมตีเชิงปรปักษ์ที่จำเพาะเจาะจงโมเดลตัวอื่น ๆ จะยังสามารถให้ผลลัพธ์ที่ถูกต้องเพื่อใช้ในการตรวจสอบและยืนยันซึ่งกันและกันได้ เป็นการป้องกันความล้มเหลวที่จุดเดี่ยว (Single Point of Failure)

2.6) ความสามารถในการฟื้นตัว

- 2.6.1) หลักการการออกแบบให้ระบบสามารถ กู้คืนกลับสู่สภาวะปกติที่เชื่อถือได้ หลังจากเกิดเหตุการณ์ด้านความมั่นคงปลอดภัย เช่น ข้อมูลหรือโมเดลถูกบิดเบือนหรือทำลาย
- 2.6.2) ตัวอย่างในสถาปัตยกรรมปัญญาประดิษฐ์ จัดทำกระบวนการ สำรองข้อมูล ที่ครอบคลุมทั้ง ชุดข้อมูลที่ใช้ฝึกสอน และโมเดลที่ผ่านการตรวจสอบแล้วอย่างสม่ำเสมอ ในกรณีที่ตรวจพบว่าข้อมูลหรือโมเดลถูกโจมตีจนเสียหาย เช่น จากการโจมตีแบบ Data Poisoning ขั้นรุนแรง สถาปัตยกรรมต้องรองรับ การกู้คืนระบบกลับสู่สถานะที่เชื่อถือได้ล่าสุดได้อย่างรวดเร็ว เพื่อลดผลกระทบต่อการดำเนินงาน

ระยะที่ 1: ออกแบบอย่างมั่นคงปลอดภัย (Secure Design)

3) ข้อพิจารณาเพิ่มเติมสำหรับระบบปัญญาประดิษฐ์ที่จัดหาจากภายนอก

ในกรณีที่องค์กรจัดหาหรือใช้บริการแพลตฟอร์มระบบปัญญาประดิษฐ์จากผู้ให้บริการภายนอก ซึ่งองค์กรไม่ได้เป็นผู้พัฒนาโดยตรง การดำเนินการจะเปลี่ยนจากการออกแบบสถาปัตยกรรมเอง ไปเน้นที่กระบวนการกำกับดูแลและทวนสอบความน่าเชื่อถือของผู้ให้บริการ โดยมีกิจกรรมที่ต้องพิจารณาดังนี้

3.1) การประเมินความน่าเชื่อถือของผู้ให้บริการ

- 3.1.1) ทวนสอบว่าผู้ให้บริการได้รับการรับรองตามมาตรฐานสากลหรือไม่ เช่น ISO/IEC 27001:2022 ISO/IEC 42001:2023 หรือ SOC 2
- 3.1.2) ร้องขอเอกสารที่เกี่ยวข้อง เช่น รายงานผลการทดสอบเจาะระบบหรือนโยบายการจัดการความมั่นคงปลอดภัยของระบบปัญญาประดิษฐ์ของผู้ให้บริการ

3.2) การสร้างแบบจำลองภัยคุกคามในส่วนที่เชื่อมต่อ

- 3.2.1) แม้จะไม่สามารถสร้างแบบจำลองภัยคุกคามของตัวแพลตฟอร์มได้ แต่องค์กร ต้องสร้างแบบจำลองภัยคุกคามในส่วนของการเชื่อมต่อระหว่างระบบขององค์กรและผู้ให้บริการเพื่อประเมินความเสี่ยง เช่น การรั่วไหลของข้อมูลระหว่างการส่งผ่านหรือการใช้ API key ในทางที่ผิด

3.3) ข้อตกลงทางสัญญาและกฎหมาย

- 3.3.1) ตรวจสอบให้แน่ใจว่าสัญญาระบุถึงความรับผิดชอบของผู้ให้บริการอย่างชัดเจน กรณีเกิดเหตุการณ์ด้านความมั่นคงปลอดภัย
- 3.3.2) ทบทวนนโยบายความเป็นส่วนตัวของผู้ให้บริการ เพื่อทำความเข้าใจว่าข้อมูลที่องค์กรส่งไปจะถูกนำไปใช้เพื่อวัตถุประสงค์อื่น เช่น การนำไปฝึกสอนโมเดลต่อ หรือไม่ และข้อมูลถูกจัดเก็บในเขตอำนาจกฎหมายใด

ระยะที่ 1: ออกแบบอย่างมั่นคงปลอดภัย (Secure Design)

4. ผลลัพธ์ที่คาดหวัง (Expected Outcomes)

เมื่อสิ้นสุดระยะที่ 1 องค์กรควรมีเอกสารประกอบการออกแบบที่ชัดเจน ได้แก่

4.1) เอกสารแบบจำลองภัยคุกคาม ระบุภัยคุกคาม ช่องโหว่ และมาตรการควบคุมที่จำเป็นสำหรับระบบปัญญาประดิษฐ์

4.2) แผนภาพและข้อกำหนดสถาปัตยกรรมที่มั่นคงปลอดภัย แสดงโครงสร้างของระบบและกลไกความปลอดภัยที่ออกแบบไว้

4.3) ข้อกำหนดด้านความมั่นคงปลอดภัยโดยละเอียด ซึ่งจะถูกส่งมอบให้ทีมพัฒนาเพื่อนำไปปฏิบัติในระยะที่ 2 ต่อไป

ระยะที่ 2: พัฒนาอย่างมั่นคงปลอดภัย (Secure Development)

- ระยะการพัฒนาอย่างมั่นคงปลอดภัยคือ ขั้นตอนการแปรเปลี่ยนพิมพ์เขียวเชิงทฤษฎีและข้อกำหนดด้านความปลอดภัยจากระยะการออกแบบในระยะที่ 1 ให้กลายเป็นผลลัพธ์เชิงปฏิบัติที่มั่นคงปลอดภัย
- วัตถุประสงค์หลักของระยะนี้คือการนำสถาปัตยกรรมที่ออกแบบไว้มาสร้างเป็นโคด ฝึกสอนโมเดล และประกอบส่วนต่าง ๆ ขึ้นเป็นระบบต้นแบบ โดยยังคงรักษาและบังคับใช้มาตรการควบคุมความมั่นคงปลอดภัยอย่างเคร่งครัดตลอดกระบวนการ
- กิจกรรมหลักในระยะนี้มุ่งเน้นไปที่ความมั่นคงปลอดภัยของห่วงโซ่อุปทาน การคุ้มครองสินทรัพย์ระหว่างการพัฒนา และการประยุกต์ใช้แนวปฏิบัติการเขียนโคดที่ปลอดภัย

ระยะที่ 2: พัฒนาอย่างมั่นคงปลอดภัย (Secure Development)

1. การบริหารจัดการความมั่นคงปลอดภัยของห่วงโซ่อุปทานปัญญาประดิษฐ์ (AI Supply Chain Security Management)

1.1) การประเมินส่วนประกอบโอเพนซอร์ส

- 1.1.1) บริบทของปัญญาประดิษฐ์ การพัฒนาปัญญาประดิษฐ์พึ่งพาไลบรารีโอเพนซอร์สจำนวนมาก เช่น TensorFlow PyTorch Hugging Face Transformers ช่องโหว่ในไลบรารีเหล่านี้ เช่น ช่องโหว่ในฟังก์ชันการโหลดข้อมูลหรือการประมวลผลโมเดล สามารถถูกใช้เป็นช่องทางโจมตีระบบได้โดยตรง
- 1.1.2) มาตรการควบคุม
  - ใช้เครื่องมือ Software Composition Analysis (SCA) เพื่อสแกนหาช่องโหว่ (CVEs) ที่เป็นที่รู้จักในไลบรารีและส่วนประกอบที่เกี่ยวข้องทั้งหมด
  - จัดให้มีความสามารถในการจำแนกและติดตามส่วนประกอบซอฟต์แวร์ทั้งหมดที่ใช้ในระบบ (AI Asset Inventory) เพื่อให้สามารถบริหารจัดการและตอบสนองต่อช่องโหว่ที่ถูกค้นพบใหม่ได้อย่างมีประสิทธิภาพ โดยกระบวนการดังกล่าวควรสร้างผลลัพธ์ที่ มีโครงสร้างชัดเจนและเป็นที่ยอมรับ เพื่อให้สามารถนำไปใช้แลกเปลี่ยนข้อมูลระหว่างเครื่องมือและระบบต่าง ๆ ได้ เพื่อให้สามารถทำงานร่วมกันได้ เช่น รูปแบบที่สอดคล้องกับแนวทางของ ISO/IEC 5962 (SPDX) หรือ ECMA-424 (CycloneDX) เป็นต้น หรือแนวทางที่องค์กรได้นำมากำหนดและประยุกต์ใช้ในองค์กร

ระยะที่ 2: พัฒนาอย่างมั่นคงปลอดภัย (Secure Development)

1. การบริหารจัดการความมั่นคงปลอดภัยของห่วงโซ่อุปทานปัญญาประดิษฐ์ (AI Supply Chain Security Management)

1.2) การทวนสอบโมเดลที่ฝึกไว้ล่วงหน้า

- 1.2.1) บริบทของปัญญาประดิษฐ์ การใช้โมเดลที่ฝึกไว้ล่วงหน้าจากแหล่งภายนอก เช่น Model Zoos มีความเสี่ยงที่โมเดลเหล่านั้นอาจถูกฝัง Backdoor หรือมัลแวร์ โดยใช้เทคนิค Model Poisoning ซึ่งจะทำงานเมื่อได้รับข้อมูลนำเข้าที่จำเพาะเจาะจง
- 1.3) มาตรการควบคุม
  - 1.3.1) ดาวน์โหลดโมเดลจากแหล่งที่น่าเชื่อถือและผ่านการรับรองเท่านั้น
  - 1.3.2) ตรวจสอบค่าผลรวมตรวจสอบ (Checksum) ของไฟล์โมเดลเพื่อให้แน่ใจว่าไม่มีการเปลี่ยนแปลง
  - 1.3.3) ใช้เครื่องมือสแกนโมเดลเพื่อตรวจหาสิ่งผิดปกติหรือโคดอันตรายที่อาจแฝงอยู่ก่อนนำมาใช้งาน

ระยะที่ 2: พัฒนาอย่างมั่นคงปลอดภัย (Secure Development)

2. การคุ้มครองสินทรัพย์ในกระบวนการพัฒนา (Asset Protection during Development)

สินทรัพย์ของปัญญาประดิษฐ์ ไม่ได้จำกัดอยู่แค่ซอร์สโคด แต่ยังรวมถึงองค์ประกอบอื่นๆ ที่มีความสำคัญและมูลค่าสูง ซึ่งต้องได้รับการคุ้มครองอย่างเหมาะสมตลอดกระบวนการพัฒนา (อ้างอิง ISO/IEC 27002:2022 #Asset_management)

2.1) ชุดข้อมูลฝึกสอนและข้อมูลทดสอบ

- 2.1.1) ความเสี่ยง การรั่วไหลของข้อมูลฝึกสอนอาจละเมิดความเป็นส่วนตัวของเจ้าของข้อมูลอย่างรุนแรง
- 2.1.2) มาตรการควบคุม บังคับใช้นโยบาย การเข้ารหัสลับข้อมูลขณะจัดเก็บ กำหนดสิทธิ์การเข้าถึงข้อมูลอย่างรัดกุมด้วยหลักการกำหนดสิทธิ์เข้าถึงให้น้อยที่สุด และใช้เทคนิคการปิดข้อมูล หรือการทำข้อมูลนิรนาม กับข้อมูลในสภาพแวดล้อมที่ไม่ใช่การใช้งานจริง (อ้างอิง OWASP #DEVDATAPROTECT)

ระยะที่ 2: พัฒนาอย่างมั่นคงปลอดภัย (Secure Development)

2. การคุ้มครองสินทรัพย์ในกระบวนการพัฒนา (Asset Protection during Development)

2.2) โมเดลปัญญาประดิษฐ์

- 2.2.1) ความเสี่ยง โมเดลที่ฝึกสอนเสร็จแล้วเป็นทรัพย์สินทางปัญญาที่มีมูลค่าสูงและอาจเป็นเป้าหมายของการลักลอบนำออกไป
- 2.2.2) มาตรการควบคุม จัดเก็บโมเดลในที่ปลอดภัยและมีการควบคุมการเข้าถึงที่เข้มงวด ใช้ระบบควบคุมเวอร์ชั่นสำหรับโมเดล เช่น Data Version Control (DVC) หรือ Git Large File Storage (LFS) เพื่อติดตามการเปลี่ยนแปลงและสามารถกู้คืนเวอร์ชั่นที่ปลอดภัยได้ และมีมาตรการปกป้องที่เหมาะสม เช่น การเข้ารหัสลับ โมเดลเมื่อไม่ได้ใช้งาน

2.3) การควบคุมเวอร์ชั่นของสินทรัพย์ปัญญาประดิษฐ์

- 2.3.1) ความเสี่ยง: การเปลี่ยนแปลงของชุดข้อมูลฝึกสอนหรือโมเดลโดยไม่มีการควบคุมเวอร์ชั่น ทำให้ไม่สามารถตรวจสอบย้อนกลับ หรือกู้คืนเวอร์ชั่นที่ปลอดภัยได้เมื่อเกิดปัญหามาตรการควบคุม
- 2.3.2) ซอร์สโคด: ใช้ระบบควบคุมเวอร์ชั่นมาตรฐาน เช่น Git
- 2.3.3) ชุดข้อมูลและโมเดล: ใช้เครื่องมือที่ออกแบบมาเพื่อควบคุมเวอร์ชั่นของไฟล์ขนาดใหญ่และชุดข้อมูลโดยเฉพาะ เช่น DVC หรือ Git LFS เพื่อให้สามารถติดตามการเปลี่ยนแปลงของข้อมูลที่ใช้ฝึกสอนและโมเดลที่ได้ในแต่ละเวอร์ชั่นได้อย่างเป็นระบบ

2.4) Prompt และไฟล์การกำหนดค่า

- 2.4.1) ความเสี่ยงสำหรับ LLM System Prompts หรือชุดคำสั่งที่ออกแบบมาอย่างดีถือเป็นความลับทางการค้า การรั่วไหลอาจทำให้คู่แข่งลอกเลียนแบบความสามารถของระบบได้
- 2.4.2) มาตรการควบคุม จัดการ Prompts และ API Keys เสมือนเป็นข้อมูลลับ จัดเก็บในระบบจัดการข้อมูลลับโดยเฉพาะ และห้ามเก็บเป็นข้อความธรรมดา (Plain Text) ในซอร์สโคดเด็ดขาด

ระยะที่ 2: พัฒนาอย่างมั่นคงปลอดภัย (Secure Development)

3. แนวปฏิบัติการเขียนโคดและการแบ่งแยกสภาพแวดล้อม (Secure Coding and Environment Segregation)

3.1) การเขียนโคดอย่างปลอดภัยสำหรับแอปพลิเคชันปัญญาประดิษฐ์

- 3.1.1) บริบทของปัญญาประดิษฐ์ นอกเหนือจากช่องโหว่ทั่วไป โคดที่ใช้ในแอปพลิเคชันปัญญาประดิษฐ์ต้องจัดการกับความเสี่ยงเฉพาะทาง
- 3.1.2) มาตรการควบคุม
    - การตรวจสอบความถูกต้องของข้อมูลนำเข้า โคดที่รับ Prompt จากผู้ใช้ต้องผ่านการตรวจสอบและคัดกรองอย่างเข้มงวด เพื่อป้องกันการโจมตีแบบ Prompt Injection
    - การจัดการข้อผิดพลาดอย่างปลอดภัย หลีกเลี่ยงการแสดงข้อมูลทางเทคนิคที่ละเอียดอ่อนเกี่ยวกับสถาปัตยกรรมของโมเดลหรือไลบรารีที่ใช้ ซึ่งอาจเป็นประโยชน์ต่อผู้โจมตี
    - ใช้เครื่องมือ Static Application Security Testing (SAST) สแกนหาช่องโหว่ในโคดที่เขียนขึ้นเองอย่างสม่ำเสมอ (อ้างอิง ISO/IEC 27002 #Application_security)

ระยะที่ 2: พัฒนาอย่างมั่นคงปลอดภัย (Secure Development)

3. แนวปฏิบัติการเขียนโคดและการแบ่งแยกสภาพแวดล้อม (Secure Coding and Environment Segregation)

3.2) การแบ่งแยกสภาพแวดล้อม

- 3.2.1.) บริบทของปัญญาประดิษฐ์ การแยกระหว่างสภาพแวดล้อมการพัฒนาการทดสอบ และการใช้งานจริง เป็นสิ่งจำเป็นอย่างยิ่งเพื่อจำกัดขอบเขตความเสียหาย โดยแต่ละสภาพแวดล้อมมีวัตถุประสงค์ที่แตกต่างกันดังนี้
  - **การพัฒนา** เป็นสภาพแวดล้อมที่เปลี่ยนแปลงบ่อยที่สุดสำหรับนักพัฒนาในการ
  - เขียนโคดและทดสอบเบื้องต้น ไม่มีความเสถียรและไม่เหมือนกับสภาพแวดล้อมจริง
  - **การทดสอบ** เป็นสภาพแวดล้อมที่จำลองให้ใกล้เคียงกับการใช้งานจริงมากที่สุด มีไว้เพื่อเป็น "ด่านสุดท้าย" สำหรับการทดสอบที่ครอบคลุม เช่น การทดสอบโดยผู้ใช้ (UAT) การทดสอบประสิทธิภาพ และการทดสอบเจาะระบบ เพื่อให้มั่นใจว่าระบบทำงานได้ถูกต้องและปลอดภัยก่อนนำขึ้นใช้งานจริง การมีสภาพแวดล้อมนี้ช่วยป้องกันไม่ให้ข้อผิดพลาดหลุดรอดไปสร้างความเสียหายในการใช้งานจริงได้
  - **การใช้งานจริง** เป็นสภาพแวดล้อมสำหรับให้บริการผู้ใช้จริง มีความเสถียรและต้องมีมาตรการควบคุมที่เข้มงวดที่สุด
- 3.2.2) มาตรการควบคุม สภาพแวดล้อมการใช้งานจริงต้องถูกแยกขาดและมีมาตรการควบคุมที่เข้มงวดที่สุด สภาพแวดล้อมการทดสอบควรใช้ชุดข้อมูลที่ไม่ละเอียดอ่อน เพื่อทำการทดสอบ และการเข้าถึงระหว่างสภาพแวดล้อมต่าง ๆ ต้องถูกควบคุมและตรวจสอบอย่างรัดกุม

ระยะที่ 2: พัฒนาอย่างมั่นคงปลอดภัย (Secure Development)

4. ข้อพิจารณาด้านความมั่นคงปลอดภัยในการใช้ API ของบุคคลที่สาม

การเรียกใช้โมเดลปัญญาประดิษฐ์ผ่าน API ของผู้ให้บริการภายนอก ก่อให้เกิดความเสี่ยงด้านการรั่วไหลของข้อมูลและความเป็นส่วนตัว องค์กรต้องมีมาตรการควบคุมดังนี้

4.1) การป้องกันข้อมูลรั่วไหล

- 4.1.1) หลักการลดข้อมูลให้เหลือน้อยที่สุด ส่งข้อมูลไปยัง API เท่าที่จำเป็นต่อการประมวลผลเท่านั้น
- 4.1.2) การกรองข้อมูลก่อนส่ง ใช้เทคนิคการปิดข้อมูล หรือการทำนามแฝงเพื่อลบ หรือแทนที่ข้อมูลส่วนบุคคล หรือข้อมูลที่เป็นความลับขององค์กร ก่อนที่จะส่งข้อมูลออกจากเครือข่ายขององค์กร

4.2) ความมั่นคงปลอดภัยของช่องทางสื่อสาร

- 4.2.1) ตรวจสอบให้แน่ใจว่าการเชื่อมต่อทั้งหมดใช้โปรโตคอลการเข้ารหัสลับที่แข็งแกร่ง เช่น TLS 1.2 หรือสูงกว่า

4.3) การจัดการข้อมูลลับ

- 4.3.1) ห้าม เก็บ API Keys ในซอร์สโคดโดยเด็ดขาด แต่ให้จัดเก็บและเรียกใช้จากระบบจัดการข้อมูลลับที่ปลอดภัย

ระยะที่ 2: พัฒนาอย่างมั่นคงปลอดภัย (Secure Development)

4. ผลลัพธ์ที่คาดหวัง (Expected Outcomes)

เมื่อสิ้นสุดระยะที่ 2 องค์กรควรมีผลลัพธ์ที่พร้อมสำหรับการทวนสอบในระยะต่อไป ได้แก่

- ซอร์สโคดของแอปพลิเคชันปัญญาประดิษฐ์ที่ผ่านการตรวจสอบความปลอดภัย
- โมเดลปัญญาประดิษฐ์ที่ผ่านการฝึกสอนและจัดเก็บอย่างปลอดภัยพร้อมการควบคุมเวอร์ชั่น
- ผลลัพธ์เชิงเอกสารที่ระบุส่วนประกอบซอฟต์แวร์ทั้งหมดของระบบ (AI Asset Inventory)
- หลักฐานการกำหนดค่าความปลอดภัยและการแบ่งแยกสภาพแวดล้อม

ระยะที่ 3: ทวนสอบด้านความมั่นคงปลอดภัย (Secure Verification)

- ทวนสอบการปฏิบัติตามข้อกำหนด (Compliance Verification) — Data Governance, Cloud Config, Code Review
- ทวนสอบประสิทธิผลกลไกความปลอดภัย (Security Mechanism Verification) — Guardrails/Content Filtering, RBAC/Access Control, Alerting
- ทดสอบเชิงรุก (AI Red Teaming) — Adversarial Samples, Prompt Injection & Jailbreaking, Privacy Inference
- ผลลัพธ์: Security Test Reports, Penetration Test Report, Validated Risk Register, Go/No-Go

ระยะที่ 3: ทวนสอบด้านความมั่นคงปลอดภัย (Secure Verification)

1) การทวนสอบการปฏิบัติตามข้อกำหนด

ขั้นตอนนี้เป็นการตรวจสอบเชิงเอกสารและเชิงเทคนิคว่า ระบบที่พัฒนาขึ้นนั้นสอดคล้องกับนโยบาย ข้อกำหนด และมาตรฐานที่กำหนดไว้ในระยะการออกแบบหรือไม่ เป็นการตอบคำถามที่ว่า "เราได้สร้างสิ่งที่ตั้งใจจะสร้างหรือไม่?" (อ้างอิง ISO/IEC 27002:2022 A5.36)

1.1) บริบทของปัญญาประดิษฐ์ การทวนสอบไม่ได้จำกัดอยู่แค่โคด แต่ครอบคลุมถึงกระบวนการที่เกี่ยวข้องกับข้อมูลและโมเดลทั้งหมด

1.2) ตัวอย่างการทวนสอบ

- 1.2.1) การทวนสอบการกำกับดูแลข้อมูล ตรวจสอบสคริปต์ (Script) ที่ใช้ในการประมวลผลข้อมูล เพื่อยืนยันว่าได้มีการใช้เทคนิคการทำข้อมูลนิรนามกับข้อมูลส่วนบุคคลจริงตามที่ออกแบบไว้ และไม่มีข้อมูลที่ละเอียดอ่อนรั่วไหลไปยังชุดข้อมูลที่ใช้ในการฝึกสอน
- 1.2.2) การทวนสอบสถาปัตยกรรม ทบทวนการตั้งค่าบนระบบคลาวด์ เพื่อยืนยันว่าการแบ่งแยกสภาพแวดล้อม ระหว่าง Training และ Inference ได้ถูกนำไปใช้อย่างถูกต้อง และมี Network Policy ที่จำกัดการสื่อสารระหว่างกันจริง
- 1.2.3) การทวนสอบข้อกำหนดโคด ตรวจสอบซอร์สโคด (Code Review) เพื่อยืนยันว่ามีการเรียกใช้ไลบรารีการเข้ารหัสลับ (Encryption Library) สำหรับการจัดเก็บโมเดลตามที่ระบุไว้ในข้อกำหนดด้านความปลอดภัย

ระยะที่ 3: ทวนสอบด้านความมั่นคงปลอดภัย (Secure Verification)

2) การทวนสอบประสิทธิผลของกลไกความปลอดภัย

ขั้นตอนนี้มุ่งเน้นการทดสอบเชิงปฏิบัติเพื่อยืนยันว่ากลไกความปลอดภัยที่ติดตั้งไว้นั้น "ทำงานได้จริงตามที่คาดหวัง" เมื่อต้องเผชิญกับสถานการณ์ต่าง ๆ ไม่ใช่แค่มีอยู่ตามข้อกำหนดเท่านั้น

2.1) บริบทของปัญญาประดิษฐ์ กลไกความปลอดภัยของปัญญาประดิษฐ์มีความซับซ้อนและต้องการการทดสอบที่จำเพาะเจาะจง

2.2) ตัวอย่างการทวนสอบ

- 2.2.1) การทดสอบ Guardrails ของ LLM สร้างชุดทดสอบที่ประกอบด้วย Prompt ที่มีเนื้อหาไม่เหมาะสม เป็นอันตราย หรือพยายามชี้นำให้เกิดอคติ แล้วส่งไปยังโมเดล LLM เพื่อทวนสอบว่ากลไก Guardrails และ Content Filtering สามารถตรวจจับและปฏิเสธการตอบสนองได้อย่างถูกต้อง
- 2.2.2) การทดสอบการควบคุมการเข้าถึง พยายามเข้าถึง API สำหรับการจัดการโมเดลด้วยสิทธิ์ของผู้ใช้งานทั่วไป เพื่อยืนยันว่าระบบจะปฏิเสธการเข้าถึงตามนโยบาย Role-Based Access Control (RBAC) ที่ออกแบบไว้
- 2.2.3) การทดสอบการแจ้งเตือน จำลองการส่งคำร้องที่มีลักษณะผิดปกติ เช่น การส่งคำร้องจำนวนมากจาก IP เดียวไปยัง API ของโมเดล เพื่อทวนสอบว่าระบบเฝ้าระวัง สามารถตรวจจับและสร้างการแจ้งเตือนไปยังทีมความมั่นคงปลอดภัยได้จริง

ระยะที่ 3: ทวนสอบด้านความมั่นคงปลอดภัย (Secure Verification)

3) การทวนสอบช่องโหว่และความเสี่ยงผ่านการทดสอบเชิงรุก

ขั้นตอนนี้เป็นการทดสอบเชิงรุก หรือ "การทดสอบเชิงเจาะระบบ" โดยสมมติตัวเป็นผู้ไม่หวังดีเพื่อค้นหาช่องโหว่ที่ไม่เคยถูกพบมาก่อน เป็นการตอบคำถามที่ว่า "เรามองข้ามอะไรไปหรือไม่?" (อ้างอิง ISO/IEC 27002:2022 A8.29, A8.8 และ OWASP-AI-EXCHANGE 5. AI security testing)

3.1) บริบทของปัญญาประดิษฐ์ การทดสอบเชิงรุกสำหรับปัญญาประดิษฐ์ หรือที่เรียกว่า "AI Red Teaming" จะใช้เทคนิคการโจมตีที่จำเพาะเจาะจงกับปัญญาประดิษฐ์โดยเฉพาะ

- 3.2) ตัวอย่างการทวนสอบ
- 3.2.1) การทดสอบด้วยตัวอย่างที่เป็นปฏิปักษ์ ทีม Red Team จะสร้างตัวอย่างข้อมูลที่ผ่านการดัดแปลงอย่างแยบยล เช่น ภาพที่มี Noise เล็กน้อย ข้อความที่เปลี่ยนตัวอักษรบางตัว ซึ่งไม่เคยอยู่ในชุดข้อมูลฝึกสอน มาใช้ทดสอบโมเดลเพื่อวัดความทนทานต่อการโจมตีแบบ Evasion Attacks ในสถานการณ์จริง
- 3.2.2) การทดสอบ Prompt Injection และ Jailbreaking สำหรับ LLM ทีม Red Team จะใช้เทคนิคขั้นสูง เช่น การสวมบทบาท หรือการโจมตีทางอ้อม (Indirect Injection) เพื่อพยายามหลบเลี่ยงกลไก Guardrails ที่ป้องกันอยู่ และบังคับให้โมเดลทำงานนอกเหนือขอบเขตที่ได้รับอนุญาต
- 3.2.3) การทดสอบการอนุมานข้อมูลส่วนบุคคล ทีมทดสอบพยายามใช้ผลลัพธ์จาก API สาธารณะของโมเดล เพื่ออนุมานข้อมูลที่ละเอียดอ่อนจากชุดข้อมูลฝึกสอน เช่น การโจมตีแบบ Membership Inference เพื่อตรวจสอบว่าข้อมูลของบุคคลใดบุคคลหนึ่งถูกใช้ในการฝึกสอนหรือไม่ ซึ่งเป็นการทวนสอบความเสี่ยงด้านการรั่วไหลของข้อมูล
- 3.2.4) การทดสอบการใช้งานในทางที่ผิด ทีม Red Team จะพยายามใช้ระบบเพื่อสร้างผลลัพธ์ที่เป็นอันตราย เช่น การสร้าง Prompt เพื่อให้ปัญญาประดิษฐ์เขียนสคริปต์สำหรับมัลแวร์ หรือสร้างอีเมลฟิชชิ่งที่แนบเนียน เพื่อทวนสอบว่ากลไกป้องกันที่ออกแบบไว้สามารถทำงานได้จริง

ระยะที่ 3: ทวนสอบด้านความมั่นคงปลอดภัย (Secure Verification)

4. ผลลัพธ์ที่คาดหวัง (Expected Outcomes)

เมื่อสิ้นสุดระยะที่ 3 องค์กรควรมีเอกสารและหลักฐานที่ชัดเจนเพื่อใช้ในการตัดสินใจก่อนนำระบบขึ้นใช้งานจริง

4.1) รายงานผลการทดสอบความมั่นคงปลอดภัย สรุปผลการทวนสอบการปฏิบัติตามข้อกำหนดและผลการทดสอบประสิทธิผลของกลไกต่าง ๆ

4.2) รายงานผลการทดสอบเจาะระบบ ระบุช่องโหว่ที่ค้นพบจากการทดสอบเชิงรุก พร้อมระดับความรุนแรงและข้อเสนอแนะในการแก้ไข

4.3) ทะเบียนความเสี่ยงฉบับทบทวนที่ได้รับการปรับปรุงข้อมูลความเสี่ยงใหม่ ๆ และยืนยันว่ามาตรการควบคุมที่มีอยู่มีประสิทธิผล

4.4) การอนุมัติเพื่อนำไปใช้งาน การตัดสินใจโดยผู้มีอำนาจว่าจะอนุมัติให้นำระบบไปใช้งานจริง แก้ไขข้อบกพร่องก่อน หรือระงับโครงการ ขึ้นอยู่กับระดับความเสี่ยงที่คงเหลืออยู่

ระยะที่ 4: นำไปใช้งานอย่างมั่นคงปลอดภัย (Secure Deployment)

- เสริมความแข็งแกร่งโครงสร้างพื้นฐาน (Infrastructure Hardening) — PoLP, Hardened Images, Network Segmentation, WAF
- คุ้มครองโมเดลและข้อมูล (Model/Data Protection) — Encryption, Key Management (HSM/KMS), Confidential Computing (Trusted Execution Environment (TEE))
- ทดสอบก่อนใช้งานจริง (Pre-deployment Validation) — Config Audit, Final Red Team, Change Management
- ผลลัพธ์: ระบบ Production ที่ผ่าน Hardening, หลักฐานควบคุม, รายงานทดสอบขั้นสุดท้าย, ATO

ระยะที่ 4: นำไปใช้งานอย่างมั่นคงปลอดภัย (Secure Deployment)

1) การเสริมความแข็งแกร่งให้โครงสร้างพื้นฐาน

เป็นกระบวนการกำหนดค่าโครงสร้างพื้นฐานที่รองรับการทำงานของปัญญาประดิษฐ์ ให้มีระดับความมั่นคงปลอดภัยสูงสุด เพื่อลดพื้นผิวการโจมตี (อ้างอิง ISO/IEC 27002:2022 A8.9 Configuration management)

1.1) บริบทของปัญญาประดิษฐ์ โครงสร้างพื้นฐานของปัญญาประดิษฐ์ไม่ได้มีแค่เซิร์ฟเวอร์ แต่ยังรวมถึงบริการจัดเก็บข้อมูล Container Orchestration และ API Gateway ซึ่งแต่ละส่วนต้องได้รับการกำหนดค่าอย่างรัดกุม

1.2) ตัวอย่างมาตรการควบคุม

- 1.2.1) การบังคับใช้หลักการสิทธิ์น้อยที่สุด บัญชีที่ใช้ในการรันโมเดลปัญญาประดิษฐ์ จะต้องได้รับสิทธิ์เฉพาะที่จำเป็นต่อการทำงานเท่านั้น เช่น สิทธิ์ในการอ่านไฟล์โมเดล สิทธิ์ในการเขียนไฟล์บันทึก (Log) และสิทธิ์ในการประมวลผล แต่ต้องไม่มีสิทธิ์ในการเข้าถึงชุดข้อมูลฝึกสอน หรือแก้ไขตัวโมเดลโดยเด็ดขาด เพื่อจำกัดขอบเขตความเสียหาย หาก Inference Service ถูกบุกรุก
- 1.2.2) การใช้อิมเมจที่ผ่านการเสริมความแข็งแกร่ง ใช้ Container Image ที่ผ่านการปรับแต่งให้มีความปลอดภัย โดยการลบไลบรารี เครื่องมือ และ Shell ที่ไม่จำเป็นออกทั้งหมด เพื่อลดช่องทางที่ผู้บุกรุกจะสามารถใช้ประโยชน์ได้หลังเจาะเข้ามาในระบบ
- 1.2.3) การแบ่งส่วนเครือข่าย กำหนดค่า Network Policy ให้ API Endpoint ของโมเดลปัญญาประดิษฐ์ สามารถสื่อสารได้เฉพาะกับส่วนประกอบที่ได้รับอนุญาตเท่านั้น และป้องกันด้วย Web Application Firewall (WAF) ที่มีการตั้งค่ากฎ สำหรับป้องกันการโจมตี API โดยเฉพาะ

ระยะที่ 4: นำไปใช้งานอย่างมั่นคงปลอดภัย (Secure Deployment)

2) การคุ้มครองโมเดลและข้อมูลในสภาวะแวดล้อมการใช้งานจริง

เป็นการใช้เทคนิคการเข้ารหัสลับและเทคโนโลยีขั้นสูงเพื่อคุ้มครองสินทรัพย์ปัญญาประดิษฐ์ที่มีความสำคัญสูงสุด (โมเดลและข้อมูล) ในทุกสถานะ ทั้งขณะจัดเก็บ (At-Rest) ขณะส่งผ่าน (In-Transit) และที่สำคัญคือ ขณะประมวลผล (In-Use)

2.1) บริบทของปัญญาประดิษฐ์ โมเดลปัญญาประดิษฐ์คือทรัพย์สินทางปัญญา และข้อมูลที่ผู้ใช้ส่งเข้ามาประมวลผลอาจเป็นข้อมูลที่ละเอียดอ่อนอย่างยิ่ง

2.2) ตัวอย่างมาตรการควบคุม

- 2.2.1) การเข้ารหัสลับโมเดลและการจัดการคีย์ ไฟล์โมเดลที่จัดเก็บไว้ใน Object Storage หรือ File Server จะต้องถูกเข้ารหัสลับไว้เสมอ เมื่อระบบเริ่มต้นทำงาน จะต้องมีการเรียกใช้คีย์สำหรับถอดรหัสจากบริการจัดการคีย์โดยเฉพาะ เช่น Hardware Security Module (HSM) หรือ Key Management Service (KMS) บนคลาวด์ ซึ่งการเข้าถึงคีย์ทุกครั้งจะต้องถูกบันทึกและตรวจสอบอย่างเข้มงวด
- 2.2.2) การประมวลผลแบบเป็นความลับ เป็นเทคโนโลยีที่ใช้ฮาร์ดแวร์ในการสร้าง Trusted Execution Environment (TEE) หรือ "Secure Enclave" ซึ่งเป็นพื้นที่ที่ถูกแยกขาดและเข้ารหัสลับในหน่วยความจำ (RAM)

ประโยชน์ต่อปัญญาประดิษฐ์ เทคโนโลยีนี้ช่วยให้สามารถประมวลผลข้อมูลที่ละเอียดอ่อน เช่น ข้อมูลทางการแพทย์ บนโมเดลปัญญาประดิษฐ์ได้ โดยที่แม้แต่ผู้ให้บริการคลาวด์หรือผู้ดูแลระบบก็ไม่สามารถเข้าถึงหรือมองเห็นข้อมูลและโมเดลขณะกำลังประมวลผลได้ เป็นการยกระดับการคุ้มครองข้อมูลในขณะประมวลผล ซึ่งเป็นสิ่งจำเป็นสำหรับระบบปัญญาประดิษฐ์ ที่จัดการกับข้อมูลที่มีความอ่อนไหวสูง

ระยะที่ 4: นำไปใช้งานอย่างมั่นคงปลอดภัย (Secure Deployment)

3. การทดสอบและทวนสอบขั้นสุดท้ายก่อนการใช้งานจริง

เป็นด่านสุดท้ายในการประกันคุณภาพความมั่นคงปลอดภัย เพื่อยืนยันว่าระบบปัญญาประดิษฐ์ที่ติดตั้งบนสภาพแวดล้อมการใช้งานจริงนั้นมีความปลอดภัยตามที่คาดหวัง การทดสอบในระยะนี้มีความสำคัญเนื่องจากสภาพแวดล้อมการใช้งานจริงอาจมีการกำหนดค่าหรือนโยบายเครือข่ายที่แตกต่างจากสภาพแวดล้อม Testing (อ้างอิง ISO/IEC 42001:2023 A.6.2.5)

3.1) บริบทของปัญญาประดิษฐ์ เป็นการทวนสอบการทำงานร่วมกันของทุกองค์ประกอบในสภาพแวดล้อมจริง

ระยะที่ 4: นำไปใช้งานอย่างมั่นคงปลอดภัย (Secure Deployment)

3. การทดสอบและทวนสอบขั้นสุดท้ายก่อนการใช้งานจริง

3.2) ตัวอย่างการทวนสอบ

- 3.2.1) การตรวจสอบความถูกต้องของการตั้งค่า ใช้เครื่องมืออัตโนมัติในการสแกนและเปรียบเทียบการตั้งค่าของสภาพแวดล้อมการใช้งานจริงกับเอกสารสถานะพื้นฐานด้านความมั่นคงปลอดภัย เพื่อให้แน่ใจว่าไม่มีการตั้งค่าที่ไม่ปลอดภัยหลงเหลืออยู่
- 3.2.2) การทดสอบโดยทีม Red Team ทีม Red Team จะทำการทดสอบเจาะระบบในขอบเขตที่จำกัดบนสภาพแวดล้อม Production-Ready เพื่อทวนสอบว่ามาตรการป้องกันต่าง ๆ เช่น WAF, IAM Roles, และ Network Segmentation สามารถป้องกันการโจมตีในสถานการณ์จริงได้หรือไม่ การทดสอบในระยะนี้มีความสำคัญและควรแบ่งประเภทให้ชัดเจนตามสภาพแวดล้อม
- 3.3) การทดสอบในสภาพแวดล้อมเสมือนจริง
- 3.3.1) วัตถุประสงค์ เพื่อค้นหาช่องโหว่และความผิดพลาดอย่างเต็มรูปแบบในสภาพแวดล้อมที่ใกล้เคียงของจริงที่สุด แต่ไม่ส่งผลกระทบต่อผู้ใช้
- 3.3.2) ตัวอย่าง การทดสอบเจาะระบบ การทดสอบด้วยตัวอย่างที่เป็นปฏิปักษ์ การทดสอบภายใต้ภาระหนัก

ระยะที่ 4: นำไปใช้งานอย่างมั่นคงปลอดภัย (Secure Deployment)

3. การทดสอบและทวนสอบขั้นสุดท้ายก่อนการใช้งานจริง

3.4) การทวนสอบในสภาพแวดล้อมใช้งานจริง

- 3.4.1) วัตถุประสงค์ เพื่อยืนยันว่าการตั้งค่าและมาตรการป้องกันบนสภาพแวดล้อมจริงทำงานได้ถูกต้องตามที่ออกแบบไว้ โดยต้องเป็นการทดสอบที่ไม่ส่งผลกระทบต่อบริการ
- 3.4.2) ตัวอย่าง: การตรวจสอบความถูกต้องของการตั้งค่า การทดสอบการเชื่อมต่อเบื้องต้นหลังการติดตั้ง การทดสอบโดยทีม Red Team ในขอบเขตจำกัดเพื่อทวนสอบกลไกป้องกัน เช่น WAF หรือ IAM Roles

3.5) การทวนสอบกระบวนการนำขึ้นระบบ ตรวจสอบว่ากระบวนการนำโคดและโมเดลขึ้นสู่ การใช้งานจริง เป็นไปตามกระบวนการบริหารจัดการการเปลี่ยนแปลงที่ได้รับอนุมัติ และมีกลไกป้องกันการนำเวอร์ชันที่ไม่ผ่านการทดสอบขึ้นใช้งาน (อ้างอิง ISO/IEC 27002:2022 A8.19)

ระยะที่ 4: นำไปใช้งานอย่างมั่นคงปลอดภัย (Secure Deployment)

4. ผลลัพธ์ที่คาดหวัง (Expected Outcomes)

เมื่อสิ้นสุดระยะที่ 4 ระบบปัญญาประดิษฐ์ จะพร้อมสำหรับการให้บริการอย่างเป็นทางการ โดยมีผลลัพธ์ที่สำคัญคือ

4.1) ระบบปัญญาประดิษฐ์ ที่ทำงานบนสภาพแวดล้อมใช้งานจริงที่ผ่านการเสริมความแข็งแกร่ง

4.2) หลักฐานการติดตั้งและเปิดใช้งานมาตรการควบคุมความปลอดภัยทั้งหมด เช่น การเข้ารหัสลับ การจัดการคีย์ การตั้งค่าเครือข่าย

4.3) รายงานผลการทดสอบความปลอดภัยขั้นสุดท้ายบนสภาพแวดล้อมใช้งานจริง

4.4) การอนุมัติให้เริ่มใช้งานระบบอย่างเป็นทางการ

ระยะที่ 5: ดำเนินงานและบำรุงรักษาอย่างมั่นคงปลอดภัย (Secure Operation & Maintenance)

- บันทึกและเฝ้าระวังเชิงรุก (Proactive Logging & Monitoring) — ตรวจจับ Prompt Injection/Probing, Data/Concept Drift, พฤติกรรมโมเดลผิดปกติ
- บริหารการเปลี่ยนแปลงและอัปเดต (Secure Change & Update Management) — เวอร์ชันโมเดล, การฝึกสอนใหม่ (Retraining) ที่ควบคุมได้
- เตรียมพร้อมตอบสนองเหตุการณ์ (Incident Readiness) และการปรับปรุงต่อเนื่อง

ระยะที่ 5: ดำเนินงานและบำรุงรักษาอย่างมั่นคงปลอดภัย (Secure Operation & Maintenance)

1) การเฝ้าระวังและประเมินสถานะ

เป็นกระบวนการเฝ้าระวังระบบอย่างต่อเนื่อง เพื่อให้ทราบถึงสถานะปัจจุบัน ตรวจจับความผิดปกติ และประเมินความเสี่ยงเชิงรุก

1.1) การบันทึกข้อมูลเชิงรุกและการเฝ้าระวังพฤติกรรม

- 1.1.1) **หลักการ** วางรากฐานการเฝ้าระวังด้วยการบันทึกข้อมูลที่ละเอียดและครอบคลุมทั้งในระดับระบบข้อมูลนำเข้า ผลลัพธ์ และพฤติกรรมของโมเดล เพื่อใช้ในการวิเคราะห์เชิงรุก ตรวจจับการโจมตี และเฝ้าระวังการเสื่อมถอยของประสิทธิภาพโมเดล (อ้างอิง ISO/IEC 27002:2022 A.8.16)

ระยะที่ 5: ดำเนินงานและบำรุงรักษาอย่างมั่นคงปลอดภัย (Secure Operation & Maintenance)

1) การเฝ้าระวังและประเมินสถานะ

1.1.2) ตัวอย่าง

- **การเฝ้าระวังการโจมตี** บันทึกข้อมูลและวิเคราะห์ Prompt ทั้งหมด เพื่อตรวจจับความพยายามในการทำ Prompt Injection หรือ Jailbreaking และแจ้งเตือนทีม Security Operations Center (SOC) ทันที
- **การเฝ้าระวังการเสื่อมถอยของโมเดล** อาศัยข้อมูลที่บันทึกไว้เพื่อติดตามภาวะความเบี่ยงเบนของข้อมูล (Data Drift) และความสัมพันธ์ของข้อมูลที่เปลี่ยนไป (Concept Drift) เพื่อแจ้งเตือนให้เริ่มกระบวนการฝึกสอนโมเดลใหม่เมื่อจำเป็น ความเบี่ยงเบนของข้อมูล และความสัมพันธ์ของข้อมูลที่เปลี่ยนไปอย่างเป็นระบบ ควรใช้กระบวนการดังนี้
  - **สร้างสถานะพื้นฐาน** หลังจากนำโมเดลขึ้นใช้งานจริงให้วัดและบันทึกค่าสถิติของข้อมูลนำเข้าและตัวชี้วัดประสิทธิภาพของโมเดล เช่น ความแม่นยำ ค่าความเชื่อมั่น เพื่อใช้เป็น "สถานะปกติ"
  - **เฝ้าระวังอย่างต่อเนื่อง** ใช้เครื่องมือทางสถิติเพื่อเปรียบเทียบข้อมูลที่เข้ามาใหม่กับค่าพื้นฐาน และติดตามประสิทธิภาพของโมเดลอย่างสม่ำเสมอ
  - **กำหนดเกณฑ์แจ้งเตือน** กำหนดระดับความเบี่ยงเบนที่ยอมรับได้ หากค่าที่เฝ้าระวังเบี่ยงเบนเกินเกณฑ์ที่กำหนด ให้ระบบสร้างการแจ้งเตือนอัตโนมัติ
  - **วิเคราะห์และฝึกสอนใหม่** เมื่อได้รับการแจ้งเตือน ให้นักวิทยาศาสตร์ข้อมูลเข้าวิเคราะห์หาสาเหตุ หากยืนยันได้ว่าโมเดลเสื่อมประสิทธิภาพจริง ให้เริ่มกระบวนการฝึกสอนโมเดลใหม่ตามที่กำหนดไว้

ระยะที่ 5: ดำเนินงานและบำรุงรักษาอย่างมั่นคงปลอดภัย (Secure Operation & Maintenance)

2. การเฝ้าระวังและรับมือความเสี่ยงจากระบบปัญญาประดิษฐ์ของบุคคลที่สาม

เมื่อใช้บริการระบบปัญญาประดิษฐ์โดยผู้ให้บริการภายนอก องค์กรไม่สามารถควบคุมระบบภายในได้โดยตรง แต่สามารถบริหารความเสี่ยงได้โดย

2.1) **การเฝ้าระวังข้อมูลนำเข้าและผลลัพธ์** เฝ้าระวังข้อมูลที่นำเข้าระบบปัญญาประดิษฐ์และผลลัพธ์ที่ได้รับกลับมา เพื่อตรวจจับความผิดปกติ เนื้อหาที่ไม่เหมาะสม หรือสัญญาณการรั่วไหลของข้อมูล

2.2) **การเฝ้าระวังประสิทธิภาพบริการ** ติดตามตัวชี้วัด เช่น อัตราความผิดพลาด หรือเวลาในการตอบสนอง หากค่าเหล่านี้เปลี่ยนแปลงอย่างมีนัยสำคัญอาจเป็นสัญญาณว่าเกิดปัญหาที่ฝั่งผู้ให้บริการ

2.3) **การเตรียมแผนเผชิญเหตุ** จัดทำคู่มือปฏิบัติสำหรับกรณีที่บริการระบบปัญญาประดิษฐ์ขัดข้องหรือถูกโจมตี โดยต้องมีขั้นตอนการติดต่อผู้ให้บริการ การสลับไปใช้ระบบสำรอง (ถ้ามี) และการสื่อสารกับผู้ใช้งาน

ระยะที่ 5: ดำเนินงานและบำรุงรักษาอย่างมั่นคงปลอดภัย (Secure Operation & Maintenance)

3) การบริหารจัดการประสิทธิภาพและความพร้อมใช้งาน

3.1) **หลักการ** วางแผนสถาปัตยกรรมให้พร้อมรองรับการใช้งานที่เพิ่มขึ้นเพื่อรักษาความพร้อมใช้งานของระบบและป้องกันปัญหาระบบไม่สามารถให้บริการได้

3.2) **ตัวอย่าง**

- 3.2.1) **การทดสอบภาระงาน** จำลองการใช้งานจำนวนมากอย่างสม่ำเสมอ เพื่อหาจุดคอขวด
- 3.2.2) **การลดและขยายระบบอัตโนมัติ** ออกแบบสถาปัตยกรรมให้สามารถเพิ่มและลดทรัพยากรได้อัตโนมัติตามปริมาณการใช้งานจริง

ระยะที่ 5: ดำเนินงานและบำรุงรักษาอย่างมั่นคงปลอดภัย (Secure Operation & Maintenance)

4.) การบริหารจัดการช่องโหว่ทางเทคนิค

4.1) **หลักการ** ค้นหา ประเมิน และแก้ไขช่องโหว่ในทุกองค์ประกอบของระบบอย่างต่อเนื่อง (อ้างอิง ISO/IEC 27002:2022 A8.8)

4.2) **ตัวอย่าง**

การสแกนอัตโนมัติ ใช้เครื่องมือ SCA สแกนช่องโหว่ในไลบรารี และใช้ DAST/SAST สแกนโคดที่พัฒนาขึ้นเอง

การบริหารจัดการแพตช์ จัดทำกระบวนการประเมินความรุนแรงและอัปเดตแพตช์ความปลอดภัยตามลำดับความสำคัญ โดยต้องผ่านการทดสอบก่อนเสมอ

ระยะที่ 5: ดำเนินงานและบำรุงรักษาอย่างมั่นคงปลอดภัย (Secure Operation & Maintenance)

2) การบริหารจัดการการเปลี่ยนแปลง

เป็นกระบวนการปรับเปลี่ยนระบบอย่างมีแบบแผนและรัดกุม เพื่อให้แน่ใจว่าทุกการเปลี่ยนแปลงจะไม่นำมาซึ่งช่องโหว่ใหม่ (อ้างอิง ISO/IEC 27002:2022 8.32)

- **2.1) หลักการ** การเปลี่ยนแปลงใด ๆ ต่อระบบ ไม่ว่าจะเป็นการอัปเดตโคด การฝึกสอนโมเดลใหม่ หรือการเพิ่มความสามารถใหม่ ๆ จะต้องผ่านกระบวนการจัดการที่มั่นคงปลอดภัยและครบวงจร

ระยะที่ 5: ดำเนินงานและบำรุงรักษาอย่างมั่นคงปลอดภัย (Secure Operation & Maintenance)

2) การบริหารจัดการการเปลี่ยนแปลง

2.2) ตัวอย่างวงจรการอัปเดต

- **2.2.1) การฝึกสอนโมเดลใหม่:** เมื่อข้อมูลเปลี่ยนไปและต้องฝึกโมเดลใหม่ โมเดลเวอร์ชันใหม่จะต้องผ่านกระบวนการ ระยะที่ 3 การทวนสอบด้านความมั่นคงปลอดภัยทั้งหมดอีกครั้งก่อนนำขึ้นใช้งาน
- **2.2.2) การอัปเดตแพตช์ความปลอดภัย** เมื่อมีการออกแพตช์สำหรับไลบรารี การอัปเดตจะต้องผ่านการทดสอบการถดถอย (Regression Testing) เพื่อให้แน่ใจว่าไม่ส่งผลกระทบต่อความแม่นยำของโมเดล
- **2.2.3) การเพิ่มขีดความสามารถใหม่** เมื่อมีการร้องขอฟีเชอร์ใหม่ การพัฒนาจะต้องย้อนกลับไปสู่ระยะที่ 1 การออกแบบอย่างมั่นคงปลอดภัย เพื่อทำแบบจำลองภัยคุกคามสำหรับฟีเชอร์นั้น ๆ ก่อนดำเนินการพัฒนาและทดสอบ

ระยะที่ 5: ดำเนินงานและบำรุงรักษาอย่างมั่นคงปลอดภัย (Secure Operation & Maintenance)

3) การตอบสนองต่อเหตุการณ์และธรรมาภิบาล

เป็นกระบวนการรับมือกับเหตุการณ์ที่ไม่คาดฝัน และกำกับดูแลให้ระบบดำเนินงานภายใต้กรอบของกฎหมายและข้อบังคับ

3.1) การจัดการและการตอบสนองต่อเหตุการณ์

- **3.1.1) หลักการ** เตรียมพร้อมรับมือกับเหตุการณ์ด้านความมั่นคงปลอดภัย โดยมีแผน ทีม คู่มือปฏิบัติที่ชัดเจน และต้องมีการซ้อมแผนอย่างสม่ำเสมอ เพื่อให้แน่ใจว่าทีมสามารถตอบสนองได้อย่างมีประสิทธิภาพเมื่อเกิดเหตุการณ์จริง (อ้างอิง ISO/IEC 27002:2022 A5.26)
- **3.1.2) ตัวอย่างคู่มือปฏิบัติ** หากตรวจพบ Evasion Attack ต้องมีขั้นตอนที่ชัดเจนในการ 1) จำกัดความเสียหาย 2) กำจัด และ 3) กู้คืนระบบกลับสู่สภาวะปกติ

3.2) การกำกับดูแลและการปรับตัวตามกฎระเบียบ

- **3.2.1) หลักการ** กำกับดูแลให้ระบบปัญญาประดิษฐ์ทำงานสอดคล้องกับกฎหมาย ข้อบังคับ และนโยบายขององค์กร ซึ่งอาจเปลี่ยนแปลงได้ตลอดเวลา
- **3.2.2) ตัวอย่าง:**
  - **การตรวจสอบการปฏิบัติตามกฎระเบียบ:** ตรวจสอบระบบเป็นประจำเพื่อให้แน่ใจว่ายังคงสอดคล้องกับกฎระเบียบที่เกี่ยวข้อง เช่น กฎหมายคุ้มครองข้อมูลส่วนบุคคล หรือกฎระเบียบด้านปัญญาประดิษฐ์ใหม่ ๆ
  - **กลไกการปรับตัว:** ออกแบบระบบให้ยืดหยุ่นพอที่จะปรับเปลี่ยนได้เมื่อมีกฎหมายใหม่ประกาศใช้ เช่น รองรับสิทธิ์ในการขอลบข้อมูลออกจากชุดข้อมูลฝึกสอน

ระยะที่ 5: ดำเนินงานและบำรุงรักษาอย่างมั่นคงปลอดภัย (Secure Operation & Maintenance)

4. แหล่งข้อมูลสนับสนุนการรับมือเหตุการณ์

เพื่อให้เท่าทันภัยคุกคามรูปแบบใหม่ องค์กรควรติดตามข้อมูลจากแหล่งข่าวกรองที่เป็นที่ยอมรับในระดับสากล เช่น

4.1) หน่วยงานความมั่นคงปลอดภัยไซเบอร์ของสหภาพยุโรป (**ENISA**) ซึ่งเผยแพร่รายงานงานวิจัย และกรอบการดำเนินงานด้านความมั่นคงปลอดภัยระบบปัญญาประดิษฐ์อย่างสม่ำเสมอ

4.2) OWASP AI Exchange โครงการที่ขับเคลื่อนโดยชุมชนภายใต้ OWASP Foundation ทำหน้าที่รวบรวมเครื่องมือ ชุดข้อมูล และองค์ความรู้เกี่ยวกับภัยคุกคามและการทดสอบความปลอดภัยระบบปัญญาประดิษฐ์

ระยะที่ 5: ดำเนินงานและบำรุงรักษาอย่างมั่นคงปลอดภัย (Secure Operation & Maintenance)

5. ผลลัพธ์ที่คาดหวัง

ระยะนี้เป็นกระบวนการต่อเนื่องไม่มีจุดสิ้นสุด ตราบเท่าที่ยังคงมีการใช้งานระบบปัญญาประดิษฐ์นั้นอยู่ โดยมีผลลัพธ์ที่สำคัญคือ

5.1) ระบบเฝ้าระวังและแจ้งเตือนที่ทำงานอย่างมีประสิทธิภาพ

5.2) รายงานสรุปประสิทธิภาพ ความมั่นคงปลอดภัย และความสามารถในการรองรับภาระงานของโมเดลเป็นประจำ

5.3) บันทึกการเปลี่ยนแปลง บันทึกการจัดการช่องโหว่ทั้งหมดที่เกิดขึ้นกับระบบ รายงานสรุปเหตุการณ์ และบทเรียนที่ได้รับเพื่อนำไปปรับปรุงต่อไป

5.4) เอกสารยืนยันการปฏิบัติตามกฎระเบียบที่เป็นปัจจุบัน

5.5) การคงไว้ซึ่งสถานะความมั่นคงปลอดภัยที่แข็งแกร่งของระบบปัญญาประดิษฐ์ตลอดอายุการใช้งาน

ระยะที่ 6: กำจัดและทำลาย (Disposal)

- เลิกใช้อย่างมั่นคงปลอดภัย — ทำลายข้อมูล/โมเดลตามนโยบาย (Secure Data/Model Destruction)
- ยืนยันการปฏิบัติตามกฎหมาย/มาตรฐาน (Compliance) และบันทึกหลักฐาน (Evidence)

ระยะที่ 6: กำจัดและทำลาย (Disposal)

1) การกำจัดและทำลายข้อมูลและโมเดลอย่างมั่นคงปลอดภัย

เป็นกระบวนการทำให้สินทรัพย์ดิจิทัลทั้งหมดที่เกี่ยวข้องกับระบบปัญญาประดิษฐ์ ไม่สามารถกู้คืนหรือเข้าถึงได้อีกต่อไปอย่างถาวร ซึ่งต้องเป็นไปตามนโยบายการจัดเก็บและทำลายข้อมูลขององค์กรและข้อกฎหมายที่เกี่ยวข้อง (อ้างอิง ISO/IEC 27002:2022 A8.10 Information deletion)

1.1) บริบทของปัญญาประดิษฐ์ สินทรัพย์ดิจิทัลของปัญญาประดิษฐ์มีความหลากหลายและละเอียดอ่อน ตั้งแต่ข้อมูลฝึกสอนไปจนถึงพารามิเตอร์ของโมเดล

1.2) ตัวอย่างมาตรการควบคุม

- 1.2.1) ชุดข้อมูลฝึกสอน โดยเฉพาะข้อมูลที่มีข้อมูลส่วนบุคคล หรือความลับทางการค้า การลบไฟล์แบบปกตินั้นไม่เพียงพอ จะต้องใช้วิธีการทำลายข้อมูลอย่างปลอดภัย เช่น การเขียนทับข้อมูลซ้ำหลายครั้ง หรือการทำลายคีย์ที่ใช้เข้ารหัสลับข้อมูล ซึ่งจะทำให้ข้อมูลที่เข้ารหัสลับไว้ไม่สามารถอ่านได้อีกต่อไป
- 1.2.2) โมเดลปัญญาประดิษฐ์และ Checkpoints โมเดลที่ฝึกสอนแล้วคือ ทรัพย์สินทางปัญญาที่มีมูลค่าสูง รวมถึงไฟล์ Checkpoints ที่ถูกบันทึกไว้ระหว่างการฝึกสอนก็อาจสามารถใช้ทำวิศวกรรมย้อนกลับเพื่ออนุมานข้อมูลบางส่วนได้ สินทรัพย์เหล่านี้ทั้งหมดจะต้องถูกกำจัดด้วยวิธีการทำลายข้อมูลที่ปลอดภัยเช่นเดียวกับชุดข้อมูลฝึกสอน
- 1.2.3) ไฟล์บันทึก ไฟล์บันทึกการทำงานของระบบอาจมีข้อมูลที่ละเอียดอ่อน เช่น Prompt ที่ผู้ใช้ป้อนเข้ามา หรือผลลัพธ์ที่โมเดลสร้างขึ้น จะต้องถูกทำลายอย่างปลอดภัยตามนโยบายการเก็บรักษาข้อมูล

ระยะที่ 6: กำจัดและทำลาย (Disposal)

1.3.2) การเขียนทับข้อมูล

- **หลักการ** เขียนข้อมูลแบบสุ่มทับลงบนพื้นที่จัดเก็บข้อมูลเดิมซ้ำ ๆ หลายรอบ เพื่อให้ข้อมูลดั้งเดิมไม่สามารถกู้คืนได้
- **ข้อควรระวัง** เหมาะสำหรับฮาร์ดดิสก์แบบจานแม่เหล็ก (HDD) แต่อาจมีประสิทธิภาพลดลงหรือไม่เหมาะสมกับ Solid-State Drives (SSD) เนื่องจากเทคโนโลยีการจัดเก็บข้อมูลที่ซับซ้อนกว่า

- **1.3.3) การทำลายทางกายภาพ**
  - **หลักการ** การทำลายตัวสื่อบันทึกข้อมูลโดยตรง เป็นวิธีที่ปลอดภัยที่สุดสำหรับอุปกรณ์ที่หมดอายุการใช้งาน
  - **ตัวอย่าง** การบดย่อย (Shredding) การทำลายด้วยสนามแม่เหล็กแรงสูง (Degaussing) หรือการหลอมละลาย (Pulverizing)

ระยะที่ 6: กำจัดและทำลาย (Disposal)

2. การเรียกคืนทรัพยากรและยกเลิกสิทธิ์การเข้าถึง

เป็นกระบวนการปลดระวางโครงสร้างพื้นฐานทั้งฮาร์ดแวร์และซอฟต์แวร์ที่เคยรองรับการทำงานของระบบปัญญาประดิษฐ์ พร้อมทั้งเพิกถอนสิทธิ์การเข้าถึงและข้อมูลยืนยันตัวตนทั้งหมด เพื่อป้องกันการเกิดบัญชีผู้ใช้ร้างหรือทรัพยากรที่ไม่มีผู้ดูแลซึ่งอาจกลายเป็นช่องโหว่ได้

2.1) บริบทของปัญญาประดิษฐ์ ระบบปัญญาประดิษฐ์มักใช้ทรัพยากรประมวลผลและจัดเก็บข้อมูลขนาดใหญ่ ซึ่งต้องจัดการอย่างระมัดระวังเมื่อเลิกใช้งาน

2.2) ตัวอย่างมาตรการควบคุม

- 2.2.1) ทรัพยากรฮาร์ดแวร์และสื่อบันทึกข้อมูล เซิร์ฟเวอร์และอุปกรณ์จัดเก็บข้อมูลที่ใช้ในการฝึกสอนโมเดลปัญญาประดิษฐ์จะต้องผ่านกระบวนการล้างข้อมูลอย่างปลอดภัยก่อนนำไปใช้ใหม่ หรือหากต้องการทิ้งจะต้องทำลายทางกายภาพ เช่น การทำลายด้วยสนามแม่เหล็กหรือการบดย่อย เพื่อป้องกันการกู้คืนข้อมูล (อ้างอิง ISO/IEC 27002:2022 7.10 Storage media) ตามเทคนิคการทำลายข้อมูลที่เหมาะสม

ระยะที่ 6: กำจัดและทำลาย (Disposal)

2. การเรียกคืนทรัพยากรและยกเลิกสิทธิ์การเข้าถึง

2.2) ตัวอย่างมาตรการควบคุม

- 2.2.2) ทรัพยากรบนคลาวด์และ API Keys ดำเนินการตามแผนการปลดระวางที่กำหนดไว้ เช่น การลบ Cloud Storage Buckets ทั้งหมด การยุติการทำงานของ Virtual Machines และ Containers การลบ IAM Roles และ Service Accounts ที่สร้างขึ้นสำหรับระบบปัญญาประดิษฐ์โดยเฉพาะ และการลบ API Keys ทั้งหมดออกจากระบบจัดการข้อมูลลับ เพื่อป้องกันปัญหาการควบคุมสิทธิ์การเข้าถึงที่บกพร่อง (Broken Access Control)
- 2.2.3) บริการและ API ของบุคคลที่สาม ยกเลิกสัญญาการใช้บริการกับผู้ให้บริการข้อมูลภายนอก และเพิกถอน API keys ทั้งหมดที่ระบบปัญญาประดิษฐ์เคยใช้เชื่อมต่อ เพื่อป้องกันการใช้งานที่ไม่ได้รับอนุญาตในอนาคต

ระยะที่ 6: กำจัดและทำลาย (Disposal)

3) การปฏิบัติตามข้อกำหนดและการจัดเก็บเอกสาร

เป็นขั้นตอนสุดท้ายในเชิงบริหารจัดการ เพื่อรับประกันการปฏิบัติตามข้อกำหนดทางกฎหมายและกฎระเบียบ พร้อมทั้งเก็บรักษาหลักฐานการดำเนินงานตลอดวงจรชีวิตของระบบปัญญาประดิษฐ์ไว้เพื่อการตรวจสอบในอนาคต

3.1) บริบทของปัญญาประดิษฐ์ การมีหลักฐานที่ตรวจสอบได้เป็นสิ่งสำคัญอย่างยิ่ง โดยเฉพาะระบบปัญญาประดิษฐ์ที่ส่งผลกระทบต่อบุคคลในวงกว้าง

3.2) ตัวอย่างมาตรการควบคุม

- 3.2.1) การจัดทำใบรับรองการทำลาย สร้างและจัดเก็บเอกสารที่เป็นทางการเพื่อรับรองการทำลายข้อมูลและโมเดล เอกสารนี้ควรระบุรายละเอียดว่าสินทรัพย์ใดถูกทำลาย ใช้วิธีการใด วันที่และเวลา และผู้ที่รับผิดชอบ ซึ่งเป็นหลักฐานสำคัญสำหรับการตรวจสอบด้านการคุ้มครองข้อมูลส่วนบุคคล
- 3.2.2) การจัดเก็บเอกสารตลอดวงจรชีวิตของระบบปัญญาประดิษฐ์ เอกสารสำคัญทั้งหมดที่สร้างขึ้นตั้งแต่ระยะที่ 0 ถึง 5 เช่น รายงานการประเมินความเสี่ยง แบบจำลองภัยคุกคาม รายงานผลการทดสอบเจาะระบบ และรายงานการรับมือเหตุการณ์จะต้องถูกจัดเก็บในที่ที่ปลอดภัยตามนโยบายขององค์กร เพื่อใช้เป็นหลักฐานแสดงความรับผิดชอบ หากมีการฟ้องร้องหรือการตรวจสอบเกิดขึ้นในภายหลัง

ระยะที่ 6: กำจัดและทำลาย (Disposal)

4. ผลลัพธ์ที่คาดหวัง (Expected Outcomes)

เมื่อสิ้นสุดระยะที่ 6 ระบบปัญญาประดิษฐ์จะถูกปลดระวางอย่างสมบูรณ์และปลอดภัย โดยมีผลลัพธ์ที่สำคัญคือ

4.1) ใบรับรองการทำลายข้อมูลและโมเดล

4.2) หลักฐานการเคลียร์ทรัพยากรฮาร์ดแวร์และคลาวด์อย่างมั่นคงปลอดภัย

4.3) คลังเอกสารฉบับสมบูรณ์ของโครงการที่จัดเก็บอย่างปลอดภัย

4.4) การปิดโครงการระบบปัญญาประดิษฐ์อย่างเป็นทางการ พร้อมการยืนยันว่าความเสี่ยงคงค้างทั้งหมดได้รับการจัดการเรียบร้อยแล้ว

เอกสารอ้างอิง (Standards & References)

- ISO/IEC 22989:2022 — AI Concepts & Terminology
- ISO/IEC 27002:2022 — Secure System Architecture & Engineering Principles
- ISO/IEC 42001:2023 — AI Management System (A.5, A.6)
- ISO/IEC 23894:2023 — AI Risk Management
- OWASP AI Exchange — Development-time Threats, AI Security Testing

ข้อสรุปสำคัญ (Key Takeaways)

- บูรณาการความมั่นคงปลอดภัยตลอดวงจรชีวิต (End-to-End Security Integration)
- Shift-Left: ลงมือด้านความปลอดภัยตั้งแต่ ‘ออกแบบ’ และ ‘พัฒนา’
- วัดผลได้จริง: ผลลัพธ์/หลักฐานในแต่ละระยะ (Risk Register, AI Asset Inventory, Test Reports, Authorization to Operate (ATO))

บทที่ 4 การกำกับดูแลและการบริหารความเสี่ยงสำหรับระบบปัญญาประดิษฐ์

1. การกำกับดูแล:

กำหนดบทบาทและความรับผิดชอบ (Roles & Responsibilities)

รากฐานที่สำคัญที่สุดคือการกำหนดบทบาทและความรับผิดชอบด้านความมั่นคงปลอดภัย AI ให้ชัดเจนเป็นลายลักษณ์อักษร เพื่อป้องกันช่องว่างและความเสี่ยง

- **ผู้บริหารระดับสูง (C-Level):** กำหนดทิศทาง, อนุมัตินโยบายและทรัพยากร, และรับผิดชอบสูงสุดต่อความเสี่ยง AI ทั้งหมด.
- **เจ้าของผลิตภัณฑ์ AI (Product Owner):** รับผิดชอบความเสี่ยงทั้งหมดของระบบ AI ที่ดูแล และเป็นผู้มีอำนาจตัดสินใจ "Go/No-Go" ในการนำระบบขึ้นใช้งานจริง.
- **ทีม Data Science / ML Engineering ("ผู้สร้าง"):** พัฒนาโมเดลที่ทนทานต่อการโจมตีและปกป้องความเป็นส่วนตัวของข้อมูล.
- **ทีมความมั่นคงปลอดภัย (Security Team / "ผู้ทวนสอบ"):** ให้คำปรึกษา, ทำแบบจำลองภัยคุกคาม (Threat Modeling), และทดสอบเจาะระบบ AI (AI Red Teaming).
- **ทีมกฎหมายและ Compliance (Legal & Compliance):** ดูแลให้การพัฒนาระบบ AI สอดคล้องกับกฎหมาย เช่น PDPA และประเมินความรับผิดทางกฎหมาย.

1. การกำกับดูแล: กำหนดบทบาทและความรับผิดชอบ (Roles & Responsibilities)

ตารางที่ ๖ แสดงแผนภูมิแสดงความรับผิดชอบตามหลักการ RACI

กิจกรรม	๑. ผู้บริหารระดับสูง	๒. AI Governance Board	๓. AI Risk Committee	๔. เจ้าของระบบปัญญาประดิษฐ์ (ธุรกิจ)	๕. ทีม Data Science/ML	๖. ทีม Security	๗. ทีม Legal/ Compliance	๘. DPO	๙. ผู้ใช้งาน
๑) ด้านธรรมาภิบาลและนโยบาย									
กำหนดและอนุมัตินโยบาย/ธรรมาภิบาล AI	A	R	C	I	I	C	C	C	I
กำหนดระดับความเสี่ยงที่ยอมรับได้ (Risk Appetite)	A	C	R	C	I	C	I	I	-
จัดสรรทรัพยากร (งบประมาณ บุคลากร)	A	C	I	R	C	C	I	I	-
๒) ด้านการบริหารความเสี่ยงและกฎหมาย									
ประเมินความเสี่ยงและผลกระทบ (DPIA)	I	C	R	A	C	C	C	R	-
บูรณาการความเสี่ยง AI เข้ากับ ERM	I	I	A/R	C	I	I	I	I	-

1. การกำกับดูแล: กำหนดบทบาทและความรับผิดชอบ (Roles & Responsibilities)

กิจกรรม	๑. ผู้บริหารระดับสูง	๒. AI Governance Board	๓. AI Risk Committee	๔. เจ้าของระบบปัญญาประดิษฐ์ (ธุรกิจ)	๕. ทีม Data Science/ML	๖. ทีม Security	๗. ทีม Legal/ Compliance	๘. DPO	๙. ผู้ใช้งาน
ตรวจสอบการปฏิบัติตามกฎหมาย/ข้อบังคับ	I	C	C	R	I	I	A	R	-
๓) ด้านการพัฒนาและปฏิบัติการ									
ออกแบบและพัฒนาโมเดลที่ปลอดภัย	-	I	I	A	R	C	C	C	-
สร้างแบบจำลองภัยคุกคาม (Threat Modeling)	-	I	I	A	C	R	I	I	-
ทดสอบเจาะระบบ (AI Red Teaming)	-	I	I	A	C	R	I	I	-
อนุมัติการนำระบบขึ้นใช้งาน (Go/No-Go)	I	C	C	A	C	C	C	C	-
เฝ้าระวังและรับมือเหตุการณ์ความปลอดภัย	I	I	I	A	C	R	I	I	-
๔) ด้านการใช้งาน									
ใช้งานระบบตามนโยบาย	-	-	-	A	-	-	-	-	R
รายงานพฤติกรรมของระบบที่ผิดปกติ	-	-	I	A	C	C	-	-	R

2. การบริหารความเสี่ยง: บูรณาการ AI Risk เข้ากับทั่วทั้งองค์กร (ERM)

ความเสี่ยง AI ไม่ใช่แค่เรื่องทางเทคนิค แต่เป็น **ความเสี่ยงทางธุรกิจ** ที่ต้องถูกบูรณาการเข้ากับกรอบการบริหารความเสี่ยงระดับองค์กร (Enterprise Risk Management - ERM) เพื่อให้ผู้บริหารมองเห็นภาพรวมและตัดสินใจได้ในบริบทเดียวกับความเสี่ยงอื่นๆ

ตัวอย่างการจับคู่ความเสี่ยง AI กับความเสี่ยงทางธุรกิจ:

- **ความเสี่ยงด้านปฏิบัติการ (Operational Risk):** การโจมตีห่วงโซ่อุปทาน (AI Supply Chain Attack) เช่น การดาวน์โหลดโมเดลที่ถูกฝัง Backdoor มาใช้งาน.
- **ความเสี่ยงด้านการเงิน (Financial Risk):** การรั่วไหลของทรัพย์สินทางปัญญาผ่านการขโมยโมเดล (Model Extraction) ทำให้คู่แข่งสร้างโมเดลลอกเลียนแบบได้.
- **ความเสี่ยงด้านกฎหมาย (Compliance Risk):** การละเมิด PDPA จากการที่โมเดลเปิดเผยข้อมูลส่วนบุคคลที่ใช้ในการฝึกสอน.
- **ความเสี่ยงด้านชื่อเสียง (Reputational Risk):** การที่ AI ตัดสินใจอย่างมีอคติและไม่เป็นธรรม (เช่น คัดกรองใบสมัครงานโดยมีอคติต่อเพศ) จนสร้างความเสียหายต่อภาพลักษณ์องค์กร.

3. การปฏิบัติตามกฎหมายไทยที่สำคัญ

ระบบ AI ต้องทำงานภายใต้กฎหมายไทย โดยมี 2 ฉบับที่ส่งผลกระทบโดยตรง:

- **พ.ร.บ. การรักษาความมั่นคงปลอดภัยไซเบอร์ พ.ศ. 2562**
  - **เกี่ยวข้องเมื่อ:** ระบบ AI ถูกใช้เป็นส่วนหนึ่งของ **โครงสร้างพื้นฐานสำคัญทางสารสนเทศ (CII)** เช่น ระบบจัดการโครงข่ายไฟฟ้า หรือระบบตรวจจับการฉ้อโกงของสถาบันการเงิน.
  - **ข้อกำหนด:** ต้องมีการประเมินความเสี่ยง, มีแผนเผชิญเหตุ (Incident Response Plan) สำหรับการโจมตี AI โดยเฉพาะ และต้องมีกระบวนการบริหารจัดการการเปลี่ยนแปลง (Change Management) ที่รัดกุม

3. การปฏิบัติตามกฎหมายไทยที่สำคัญ

- **พ.ร.บ. คุ้มครองข้อมูลส่วนบุคคล พ.ศ. 2562 (PDPA):**
  - **เกี่ยวข้องใน: ทุกขั้นตอน** ของวงจรชีวิต AI ตั้งแต่การรวบรวมข้อมูลมาฝึกสอนไปจนถึงการใช้งาน
  - **ข้อกำหนด:**
    - ✓ **ต้องมีฐานทางกฎหมาย:** เช่น การขอ **"ความยินยอม" (Consent)** ที่ระบุชัดเจนว่าจะนำข้อมูลไปใช้ฝึกสอน AI
    - ✓ **ต้องมีมาตรการความปลอดภัย:** เช่น การทำนามแฝง (Pseudonymization) หรือการเข้ารหัส (Encryption) ข้อมูลที่ใช้ฝึก
    - ✓ **ต้องเคารพสิทธิเจ้าของข้อมูล:** เช่น เตรียมแนวทางสำหรับ **"สิทธิในการลบข้อมูล" (Right to Erasure)** ซึ่งอาจต้องใช้เทคนิค Machine Unlearning หรือการฝึกโมเดลใหม่ทั้งหมด

3. การปฏิบัติตามกฎหมายไทยที่สำคัญ

- **พ.ร.บ. คุ้มครองข้อมูลส่วนบุคคล พ.ศ. 2562 (PDPA):**
  - **เกี่ยวข้องใน: ทุกขั้นตอน** ของวงจรชีวิต AI ตั้งแต่การรวบรวมข้อมูลมาฝึกสอนไปจนถึงการใช้งาน
  - **ข้อกำหนด:**
    - ✓ **ต้องมีฐานทางกฎหมาย:** เช่น การขอ **"ความยินยอม" (Consent)** ที่ระบุชัดเจนว่าจะนำข้อมูลไปใช้ฝึกสอน AI
    - ✓ **ต้องมีมาตรการความปลอดภัย:** เช่น การทำนามแฝง (Pseudonymization) หรือการเข้ารหัส (Encryption) ข้อมูลที่ใช้ฝึก
    - ✓ **ต้องเคารพสิทธิเจ้าของข้อมูล:** เช่น เตรียมแนวทางสำหรับ **"สิทธิในการลบข้อมูล" (Right to Erasure)** ซึ่งอาจต้องใช้เทคนิค Machine Unlearning หรือการฝึกโมเดลใหม่ทั้งหมด

3. การปฏิบัติตามกฎหมายไทยที่สำคัญ

พระราชบัญญัติว่าด้วยการกระทำความผิดเกี่ยวกับคอมพิวเตอร์ พ.ศ. ๒๕๕๐ และฉบับแก้ไขเพิ่มเติม

- พระราชบัญญัติฉบับนี้เป็นกฎหมายหลักที่กำหนดฐานความผิดและบทลงโทษเกี่ยวกับการใช้ระบบคอมพิวเตอร์ในทางที่มิชอบ โดยมุ่งเน้นการป้องกันและปราบปรามการนำเข้าหรือเผยแพร่เนื้อหาที่ผิดกฎหมาย อันอาจส่งผลกระทบต่อความมั่นคงของประเทศและศีลธรรมอันดีของประชาชน
- การประยุกต์ใช้กับระบบปัญญาประดิษฐ์ ทั้งข้อมูลนำเข้า เช่น Prompt และข้อมูลผลลัพธ์ เช่น ข้อความ รูปภาพ วิดีโอ ถือเป็นข้อมูลคอมพิวเตอร์ภายใต้พระราชบัญญัตินี้ ดังนั้นทั้งผู้ให้บริการ (เจ้าของแพลตฟอร์ม) ที่พัฒนาหรือเปิดให้ใช้ระบบปัญญาประดิษฐ์ และผู้ใช้งานที่สร้าง Prompt เพื่อให้ระบบปัญญาประดิษฐ์สร้างเนื้อหา ต่างก็มีความรับผิดทางกฎหมายหากผลลัพธ์ที่เกิดขึ้นเข้าข่ายเป็นความผิด

3. การปฏิบัติตามกฎหมายไทยที่สำคัญ

๔) ข้อพิจารณาพิเศษ การรักษาอธิปไตยทางข้อมูลในระบบปัญญาประดิษฐ์

- นอกเหนือจากมาตรการความมั่นคงปลอดภัยเชิงเทคนิคแล้ว การกำกับดูแลระบบปัญญาประดิษฐ์จำเป็นต้องคำนึงถึงมิติของ **"อธิปไตยทางข้อมูล"** อย่างเคร่งครัด หมายถึง อำนาจในการควบคุมและบังคับใช้กฎหมายของประเทศไทยเหนือข้อมูลที่ถูกสร้าง จัดเก็บ และประมวลผลภายในประเทศ หลักการนี้
- มีความสำคัญอย่างยิ่งยวดต่อระบบปัญญาประดิษฐ์ซึ่งมีข้อมูลเป็นหัวใจสำคัญในการทำงาน เพื่อให้มั่นใจว่าการใช้ประโยชน์จากปัญญาประดิษฐ์จะไม่นำมาซึ่งความเสี่ยงต่อความมั่นคงและผลประโยชน์ของชาติ

4. การตรวจสอบและการรับรอง: กลไกสร้างความเชื่อมั่น

การตรวจสอบโดยหน่วยงานอิสระภายนอกเป็นเครื่องมือพิสูจน์ความรับผิดชอบและสร้างความไว้วางใจให้แก่ผู้มีส่วนได้ส่วนเสีย

- **ขอบเขตการตรวจสอบ AI (AI Audit Scope):**
  - **การกำกับดูแลข้อมูล (Data Governance):** ตรวจสอบแหล่งที่มาของข้อมูลและการขอความยินยอมตาม PDPA.
  - **ความเป็นธรรมและอคติ (Fairness and Bias):** ประเมินว่าโมเดลเลือกปฏิบัติหรือไม่
  - **ความทนทานและความปลอดภัย (Robustness and Security):** ทบทวนผลการทดสอบเจาะระบบ (AI Red Team)
  - **ความโปร่งใส (Transparency):** ตรวจสอบว่าระบบสามารถอธิบายเหตุผลการตัดสินใจได้หรือไม่ (Explainable AI).
- **มาตรฐานสากลที่สำคัญ:**
  - **ISO/IEC 42001:2023:** มาตรฐาน **"ระบบการจัดการปัญญาประดิษฐ์ (AIMS)"** ฉบับแรกของโลกที่สามารถขอการรับรองได้ เพื่อแสดงถึงวุฒิภาวะในการกำกับดูแล AI
  - **ISO/IEC 23894:2023:** มาตรฐาน **"คำแนะนำ"** สำหรับการบริหารความเสี่ยง AI โดยเฉพาะ ซึ่งใช้เป็นแนวทางในการประเมินความเสี่ยง

ข้อสรุปสำคัญ (Key Takeaways)

- **กำหนดผู้รับผิดชอบให้ชัดเจน (Clear Accountability)**
  - รากฐานที่สำคัญที่สุดคือการกำหนด **บทบาทและความรับผิดชอบด้าน AI** เป็นลายลักษณ์อักษร ตั้งแต่ผู้บริหารระดับสูงไปจนถึงทีมเทคนิค เพื่อไม่ให้เกิดช่องว่างและทุกคนเข้าใจหน้าที่ของตนเอง
- **มองความเสี่ยง AI ให้เป็นความเสี่ยงทางธุรกิจ (AI Risk is Business Risk)**
  - ความเสี่ยง AI ไม่ใช่แค่เรื่องเทคนิค แต่ต้องถูกผนวกเข้ากับ **กรอบการบริหารความเสี่ยงระดับองค์กร (ERM)** เพื่อให้ผู้บริหารสามารถประเมินและตัดสินใจได้ในภาพรวม เช่น การขโมยโมเดลคือ **ความเสี่ยงทางการเงิน**, AI ที่มีอคติคือ **ความเสี่ยงด้านชื่อเสียง**
- **ต้องปฏิบัติตามกฎหมายไทย (Compliance is Mandatory)**
  - ระบบ AI ต้องทำงานภายใต้ **กฎหมายไทย 2 ฉบับ** ที่สำคัญ:
    - ✓ **พ.ร.บ. ความมั่นคงปลอดภัยไซเบอร์ฯ:** เมื่อ AI ถูกใช้ในโครงสร้างพื้นฐานสำคัญ (CII)
    - ✓ **PDPA:** มีผลในทุกขั้นตอนที่เกี่ยวข้องกับข้อมูลส่วนบุคคล ตั้งแต่การเก็บข้อมูลมาฝึกสอนไปจนถึงการใช้งานจริง
- **ใช้มาตรฐานและการตรวจสอบเพื่อสร้างความเชื่อมั่น (Build Trust through Audits & Standards)**
  - การตรวจสอบโดยหน่วยงานภายนอกและการใช้ **มาตรฐานสากล** เป็นเครื่องมือพิสูจน์ว่าองค์กรมีการกำกับดูแล AI ที่ดีและมีความรับผิดชอบ โดยมี **ISO/IEC 42001** เป็นมาตรฐานหลักสำหรับระบบการจัดการ AI ที่สามารถขอการรับรองได้

ขอบพระคุณทุกท่านที่ร่วมกิจกรรม

S E C R I T Y

is not complete without U