# The Implications of AI in Privacy and Security

The Implications of AI in Privacy & Cybersecurity: Opportunities and Challenges
The Implications of AI in Privacy & Cybersecurity: Opportunities and Challenges

ศักดิ์ เสกขุนทด

*ที่ปรึกษา*

*สำนักงานพัฒนาธุรกรรมทางอิเล็กทรอนิกส์*

# AI Risks

## ความเสี่ยงที่เกี่ยวข้องกับ AI

# Risks

| Common Risk | Criminal Justice | Financial Services | Health & Social Care | Digital & Social Media | Energy & Utilities |
|---|---|---|---|---|---|
| Bias leading to discrimination | 🔴 | 🔴 | 🔴 | 🔴 | 🟡 |
| Lack of explainability | 🔴 | 🔴 | 🟡 | 🔴 | 🔴 |
| Regulator resourcing | 🟡 | 🟡 | 🔴 | 🔴 | 🔴 |
| Higher-impact cyberattacks | 🟡 | 🔴 | 🔴 | 🟡 | 🔴 |
| Failure of consent mechanisms | 🔴 | 🟡 | 🔴 | 🔴 | 🟡 |
| Loss of trust in institutions | 🟡 | 🟡 | 🔴 | 🟡 | 🔴 |
| Lack of transparency | 🔴 | 🟡 | 🟡 | 🟡 | 🟡 |
| Unequal access to services | 🟢 | 🔴 | 🔴 | 🟡 | 🟡 |
| Effects of low digital/data maturity | 🟡 | 🟡 | 🔴 | 🟢 | 🔴 |
| Erosion of privacy | 🔴 | 🟡 | 🟡 | 🟡 | 🟢 |
| Platform and data monopolies | 🟡 | 🟢 | 🟡 | 🟡 | 🔴 |
| Excessive data retention | 🔴 | 🟡 | 🟢 | 🟡 | 🟢 |
| Low 'human-in-the-loop' | 🟡 | 🟡 | 🟡 | 🟡 | 🟢 |
| Mis/disinformation | 🟢 | 🟢 | 🔴 | 🔴 | 🟢 |
| Loss of trust in AI | 🟢 | 🟢 | 🟡 | 🟢 | 🔴 |
| Undervaluation of public data | 🟡 | 🟢 | 🔴 | 🟢 | 🟢 |
| Low accuracy | 🟡 | 🟢 | 🟡 | 🟡 | 🟢 |
| Undermining professional judgement | 🟢 | 🟢 | 🔴 | 🟢 | 🟢 |
| Excessive trust in AI tools | 🟡 | 🟢 | 🟡 | 🟢 | 🟢 |

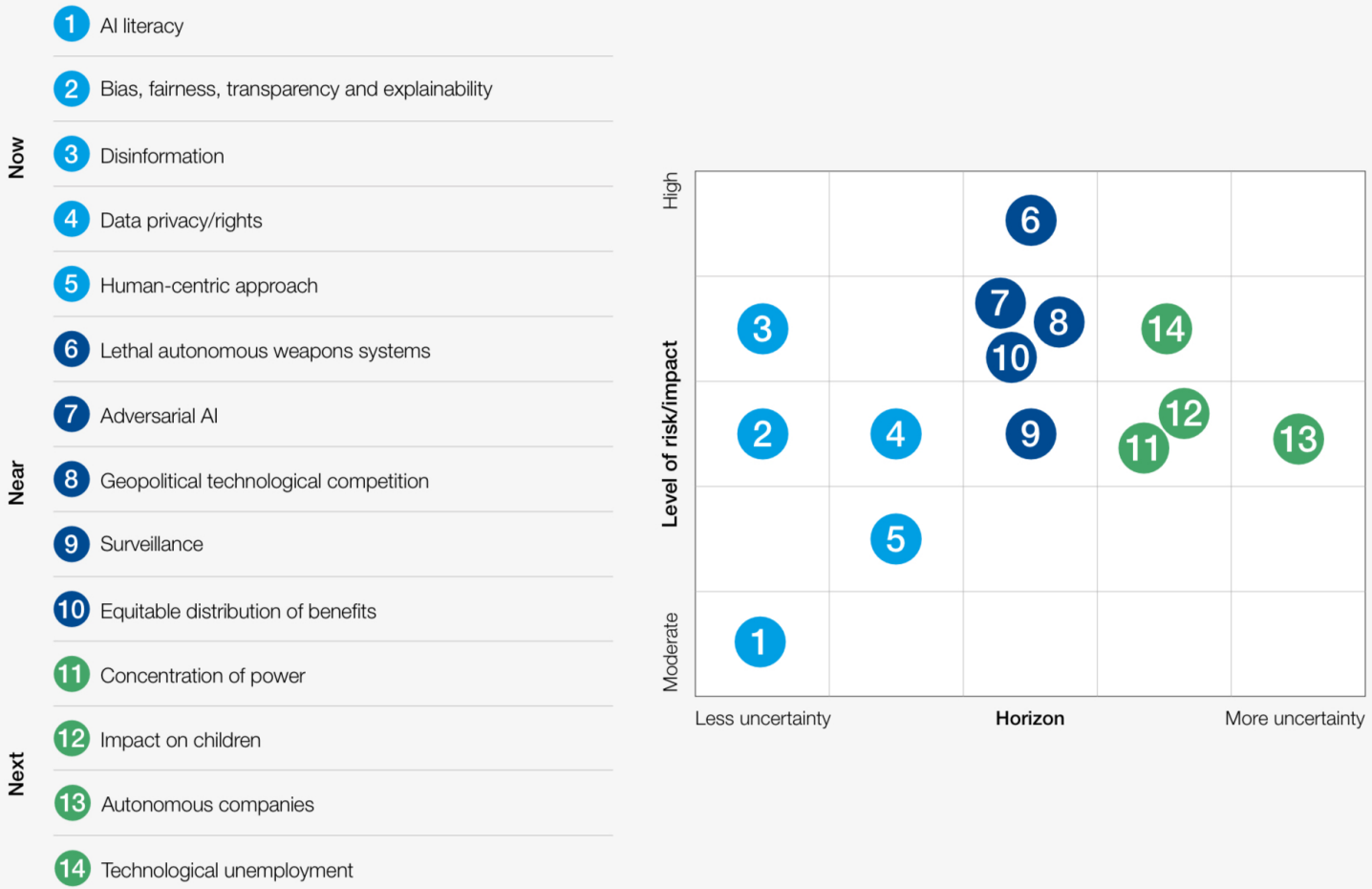
🔴 Higher Risk 🟡 Medium Risk 🟢 Lower Risk

AIGC AI GOVERNANCE CLINIC by ETDA

ADTE by ETDA

# Emerging Governance Gap of AI

FIGURE 4 | Time horizon and risk level of emerging governance gaps




**Source:** Deloitte analysis

# Risk

## ความเสี่ยง

**effect of uncertainty on objectives**

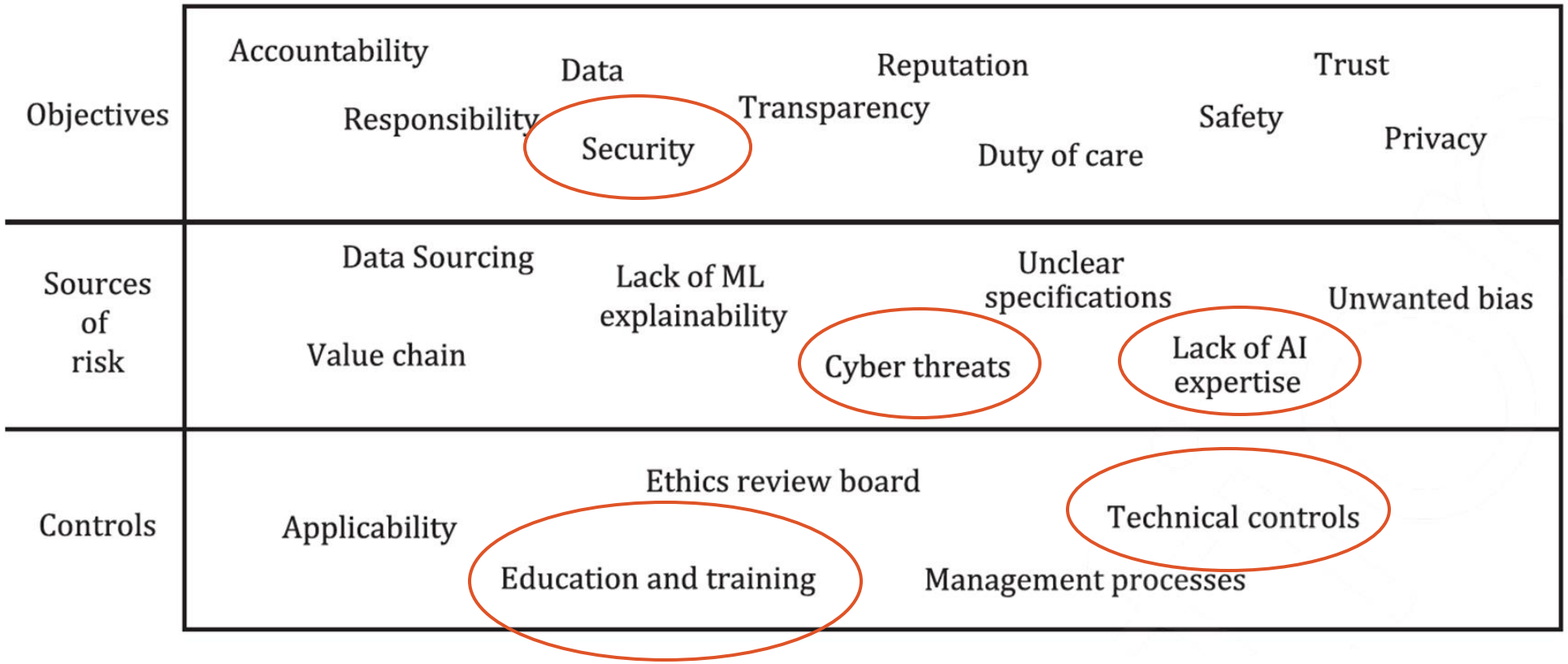
**(ผลกระทบของความไม่แน่นอนว่าจะเป็นไปตามเป้าหมายที่กำหนด)**

ISO/IEC 38507:2022 Information technology — Governance of IT — Governance implications of the use of artificial intelligence by organizations

# AI Risks

**Risk - effect of uncertainty on objectives**

**(ผลกระทบของความไม่แน่นอนว่าจะเป็นไปตามเป้าหมายที่กำหนด)**

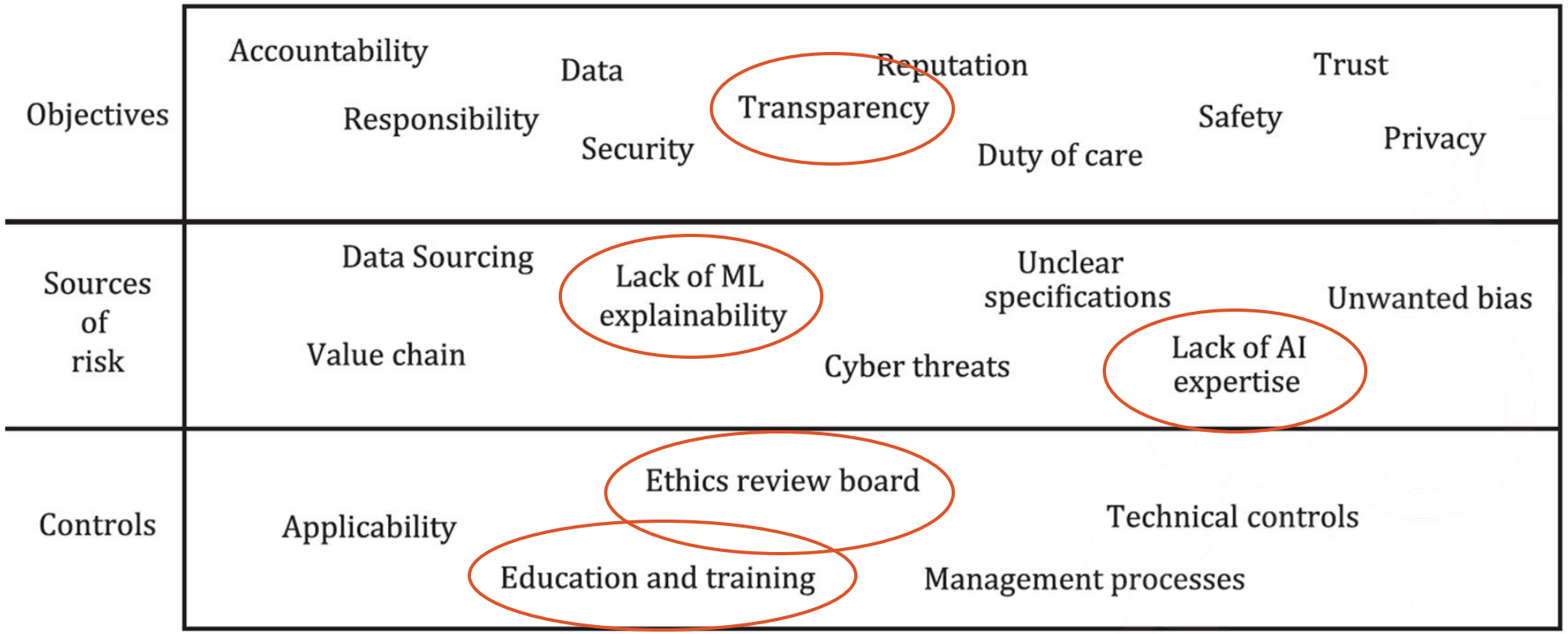



ISO/IEC 38507:2022 Information technology — Governance of IT — Governance implications of the use of artificial intelligence by organizations

# AI Risks

**Risk - effect of uncertainty on objectives**

**(ผลกระทบของความไม่แน่นอนว่าจะเป็นไปตามเป้าหมายที่กำหนด)**

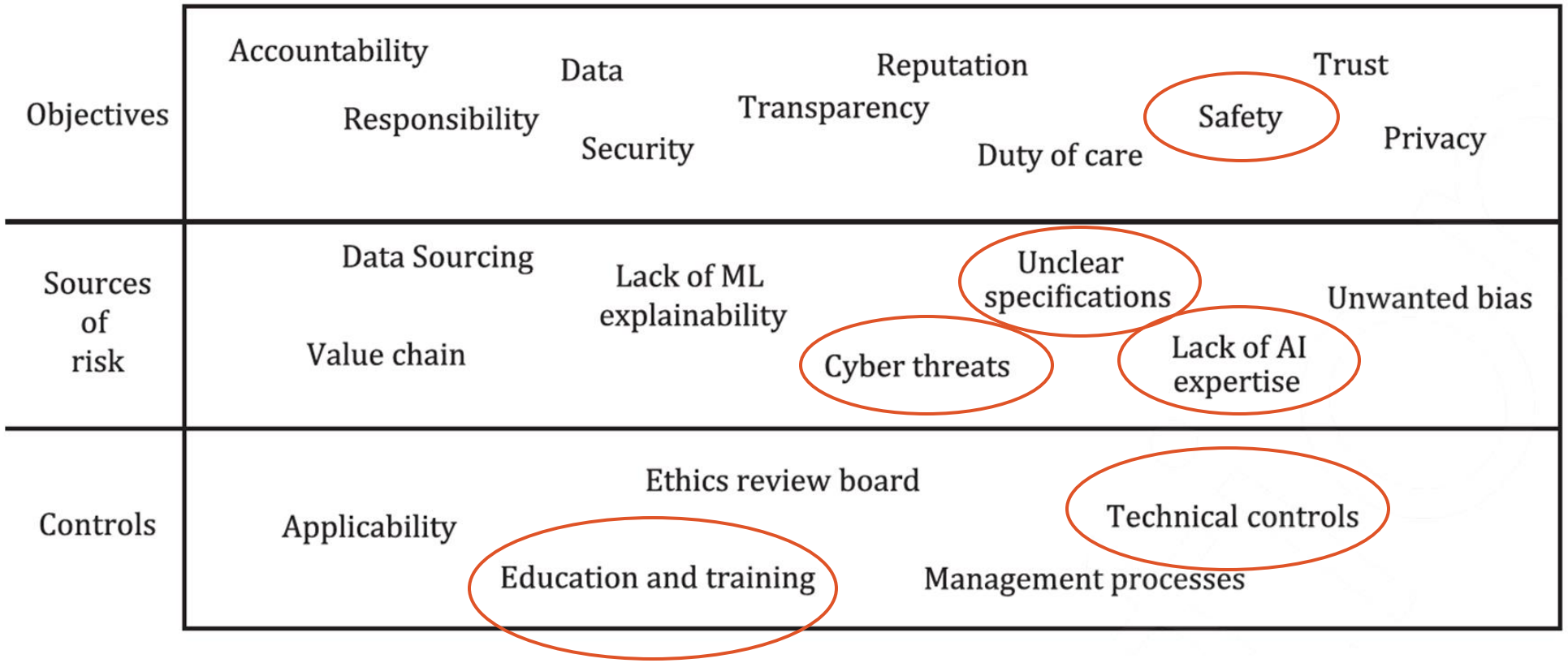



ISO/IEC 38507:2022 Information technology — Governance of IT — Governance implications of the use of artificial intelligence by organizations

# AI Risks

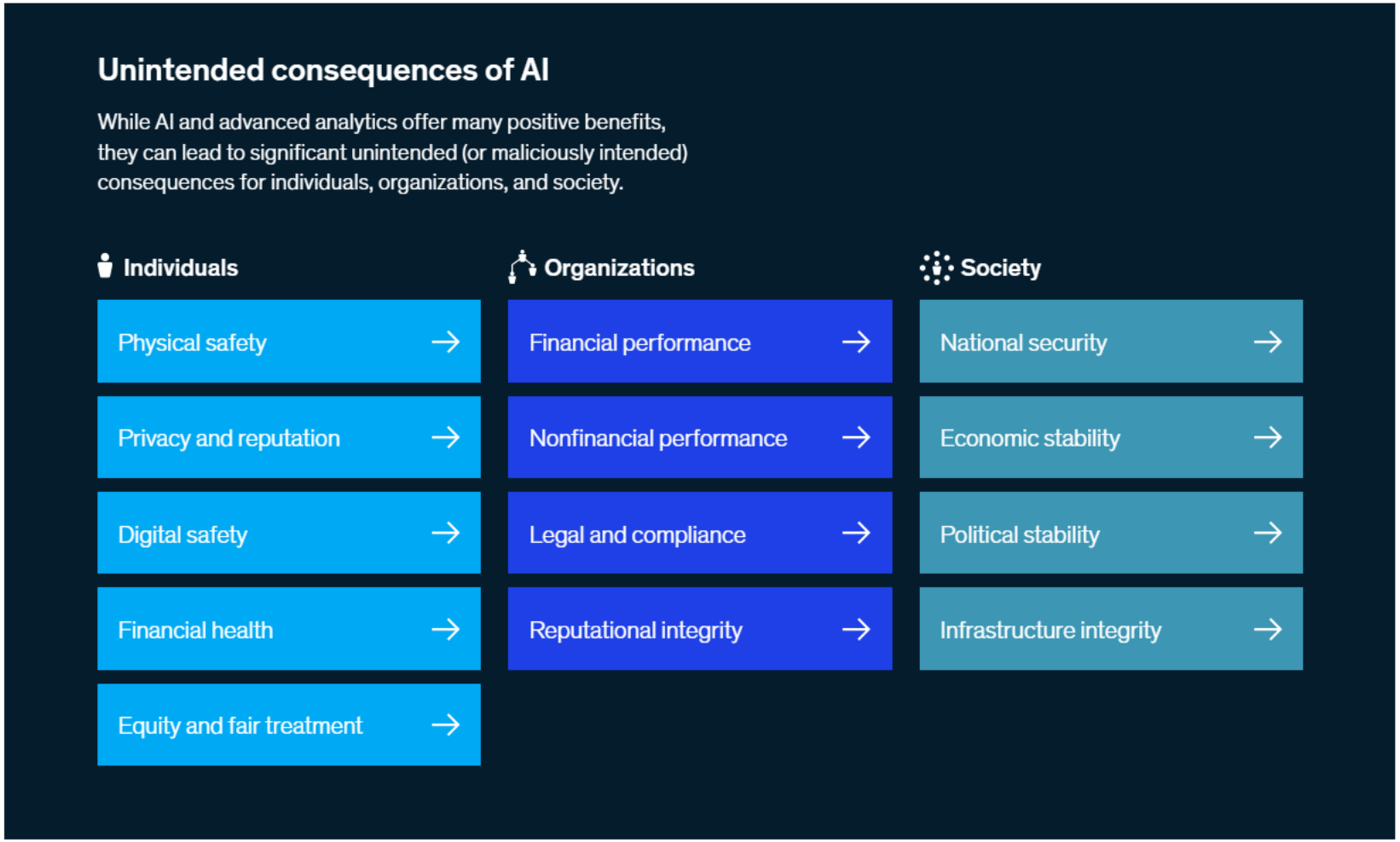
**Risk - effect of uncertainty on objectives**

**(ผลกระทบของความไม่แน่นอนว่าจะเป็นไปตามเป้าหมายที่กำหนด)**




ISO/IEC 38507:2022 Information technology — Governance of IT — Governance implications of the use of artificial intelligence by organizations

# Who could be affected and what's at risk




Source: https://www.mckinsey.com/capabilities/quantumblack/our-insights/confronting-the-risks-of-artificial-intelligence#/

# Different Context, Different Risks

**Risk – effect of uncertainty on objectives**
**(ผลกระทบของความไม่แน่นอนว่าจะเป็นไปตามเป้าหมายที่กำหนด)**

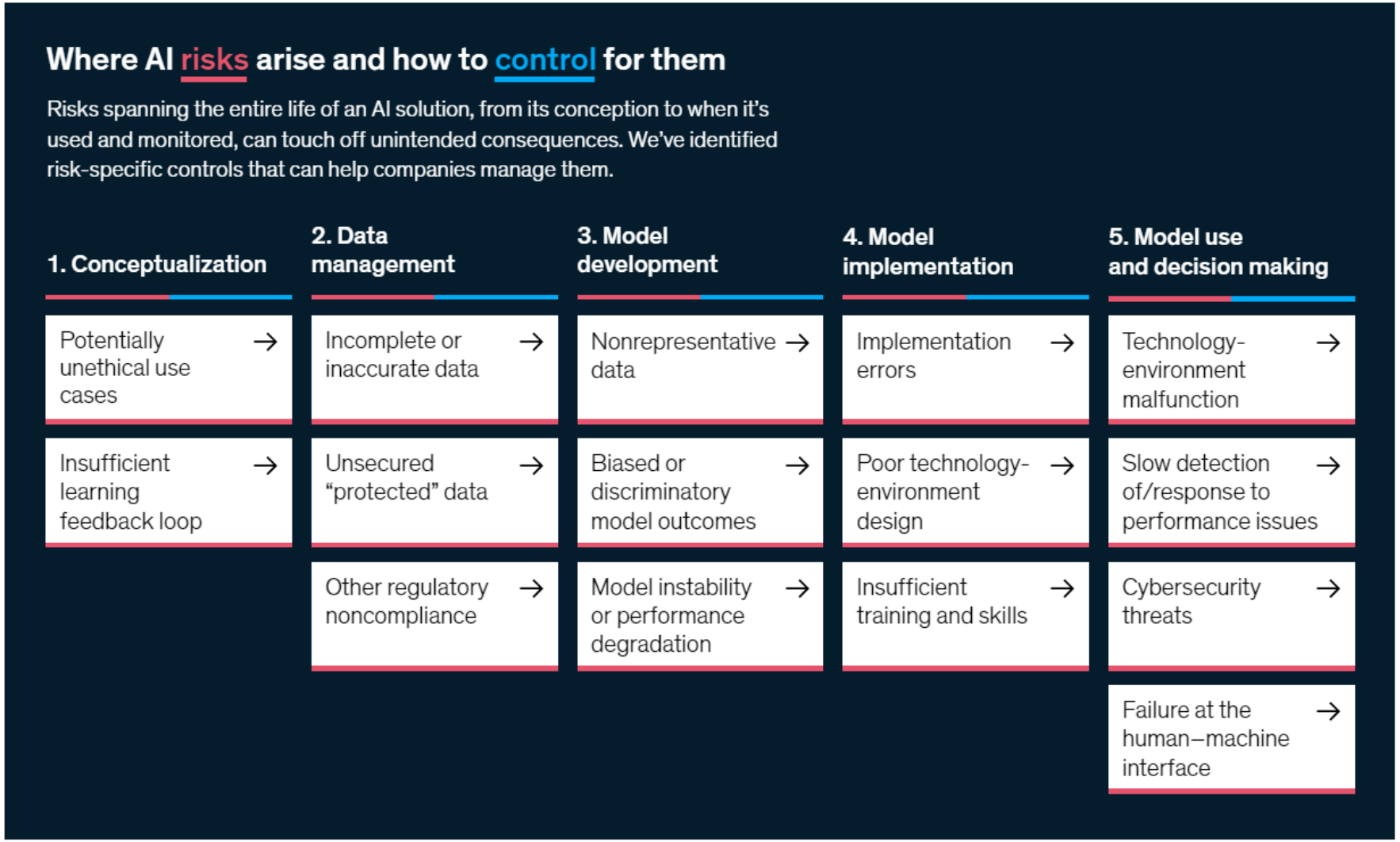
| | บริบทของการใช้งาน AI (Context) | เป้าหมาย (Objective) | ผลกระทบเชิงลบ (Negative Impact) |
|---|---|---|---|
| 1 | รถยนต์ไร้คนขับ | • ความปลอดภัย (Safety)<br>• ความมั่นคงปลอดภัย (Security)<br>• ความสอดคล้องตามกฎหมายและข้อกำหนดต่างๆ (Compliance)<br>• ความรับผิดชอบ (Accountability) | • เกิดความเสียหายต่อชีวิต ร่างกาย หรือจิตใจ<br>• ทรัพย์สินเสียหาย (ผู้เสียหายเสียทรัพย์สินเนื่องจากอุบัติเหตุ)<br>• ทรัพย์สินเสียหาย (กรณีองค์กรต้องจ่ายค่าเสียหาย)<br>• องค์กรเสียชื่อเสียง |
| 2 | การใช้ ChatGPT เพื่อตอบคำถามลูกค้า | • ความถูกต้อง (Accuracy)<br>• การคุ้มครองข้อมูลส่วนบุคคล (Privacy)<br>• ความมั่นคงปลอดภัย (Security) | • องค์กรเสียชื่อเสียง<br>• ข้อมูลส่วนบุคคลรั่วไหล<br>• ผลกระทบต่อทรัพย์สิน (ถูกปรับเนื่องจากข้อมูลส่วนบุคคลรั่วไหล)<br>• การให้บริการหยุดชะงัก |

# การประเมินความเสี่ยง

จะทำให้องค์กรเห็นถึงความไม่แน่นอนหรือ
โอกาสที่การประยุกต์ใช้ AI จะไม่เป็นไปตามเป้าหมาย
ที่กำหนด รวมถึงเห็นผลกระทบเชิงลบที่อาจเกิดขึ้น

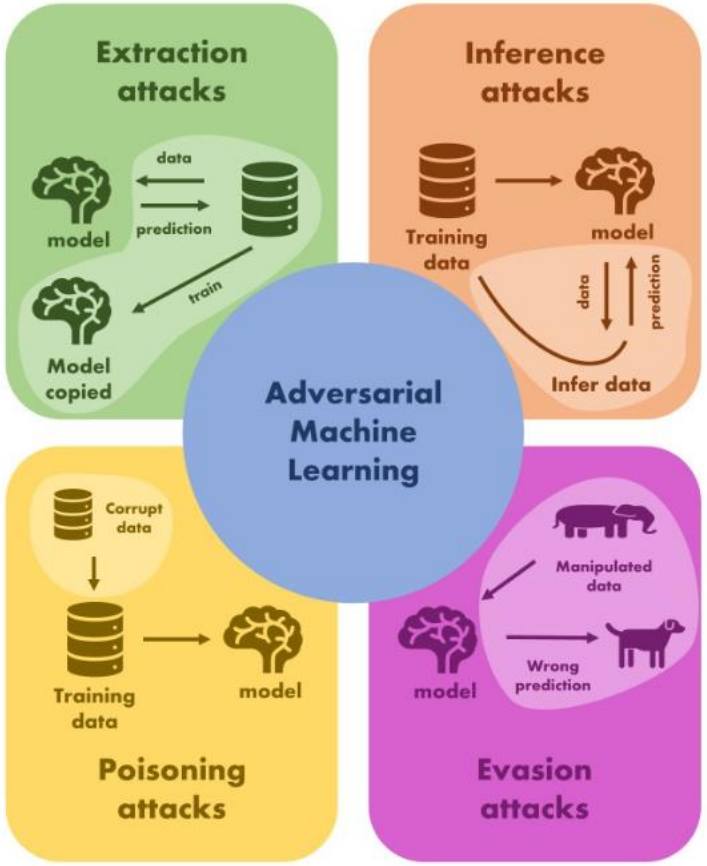
# Where AI risks arise

ความเสี่ยงสามารถเกิดได้ในทุกกระบวนการตลอดวงจรชีวิตของ AI (AI Lifecycle)




Source: https://www.mckinsey.com/capabilities/quantumblack/our-insights/confronting-the-risks-of-artificial-intelligence#/

# Adversarial Machine Learning




**Extraction Attacks:** These attacks are like a silent robbery. They aim to clandestinely extract sensitive information from a model—be it our invaluable training data or parameters.

**Inference Attacks:** Here, attackers act like detectives, inferring the output of a machine learning model by merely observing its behaviour—no direct querying required.

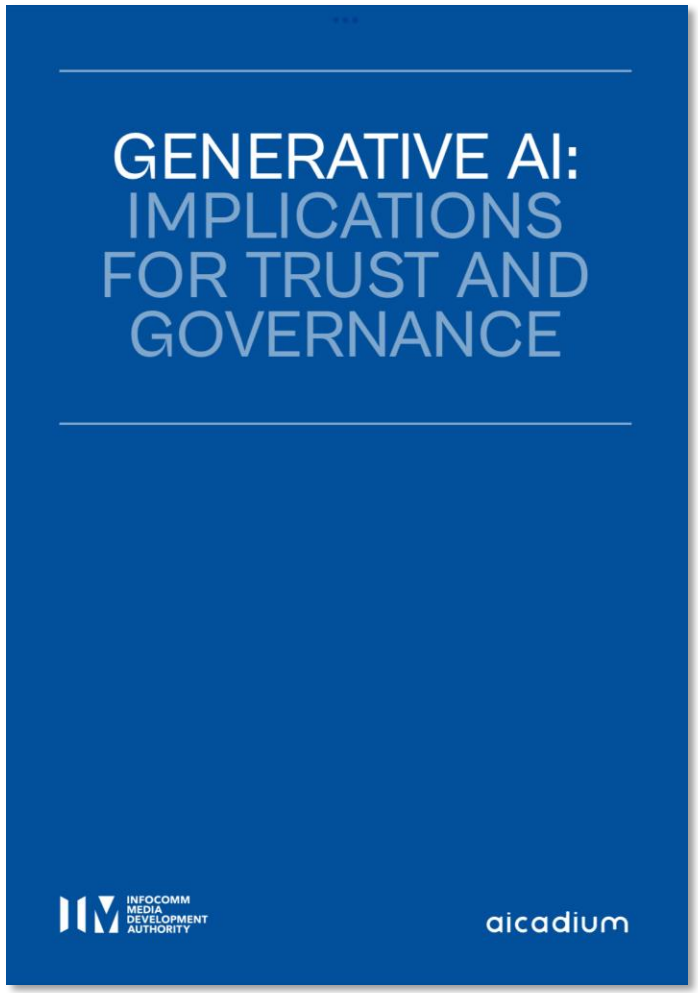
**Poisoning Attacks:** This is akin to tampering with the recipe of a dish. These attacks contaminate the training data of a machine learning model, leading it to serve incorrect predictions.

**Evasion Attacks:** In these, attackers become crafty artists, designing inputs (known as adversarial examples) that are expertly engineered to trick the model into making a wrong prediction.

Source: https://www.linkedin.com/feed/update/urn:li:activity:7098199693863624704/?utm_source=share&utm_medium=member_desktop

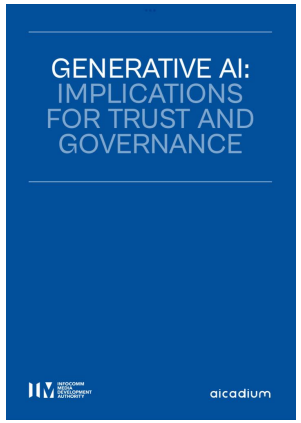
# Generative AI Risks

## ความเสี่ยงที่เกี่ยวข้องกับ Generative AI



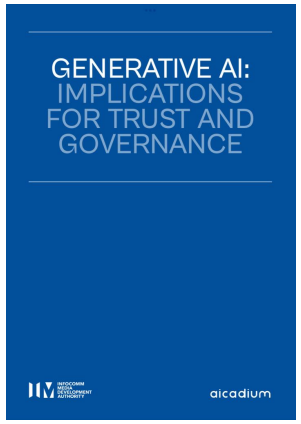

# NEW RISKS WITH GENERATIVE AI

Trustworthy AI literature has identified a few governance areas, which typically deal with robustness, explainability, algorithmic fairness, privacy and security. The Singapore Model AI Governance Framework and OECD AI Principles outline these core areas. Even though these governance areas continue to remain relevant, generative AI also poses emerging risks that may require new approaches to its governance.




# RISK 1: MISTAKES AND "HALLUCINATIONS"

Like all AI models, generative AI models make mistakes. **When generative AI makes mistakes, they are often vivid and take on anthropomorphisation, commonly known as "hallucinations".** Current and past versions of ChatGPT are known to make factual errors. Such models also have a more challenging time doing tasks like logic, mathematics, and common sense[3]. This is because ChatGPT is a model of how people use language. While language often mirrors the world, these systems however do not (yet) have a deep understanding about how the world works. Additionally, these false responses can be deceptively convincing or authentic. Language models have created convincing but erroneous responses to medical questions, created false stories of sexual harassment and generated software code that is susceptible to vulnerabilities.

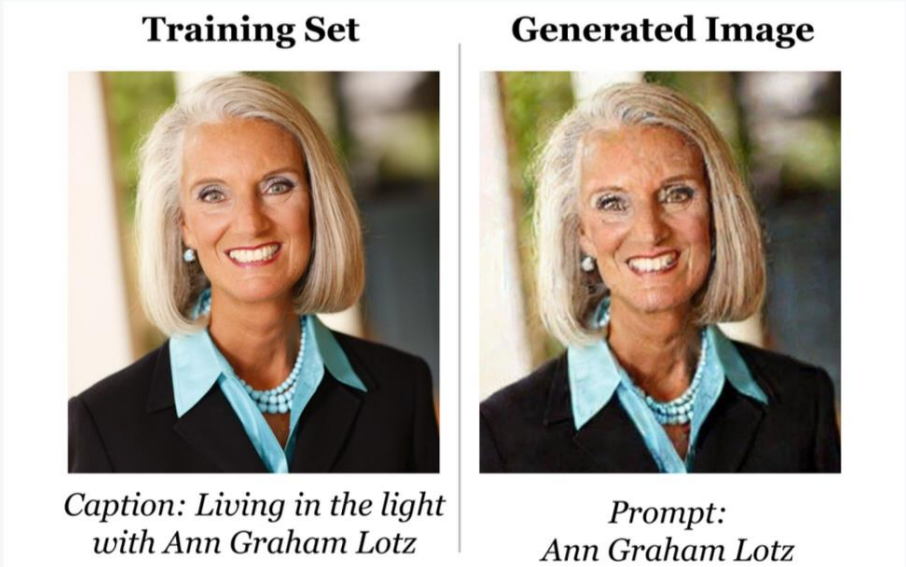



# RISK 2: PRIVACY AND CONFIDENTIALITY

Generative AI tends to have a property of "memorisation". Typically, one would expect AI models to generalise from the individual data points used to train the model, so when you use the AI model there is no trace of the underlying training data. **As the neural networks underpinning generative AI models expand, these models have a tendency to memorise**. For example, Stable Diffusion tends to memorise twice as much as older generative AI models such as GANs.

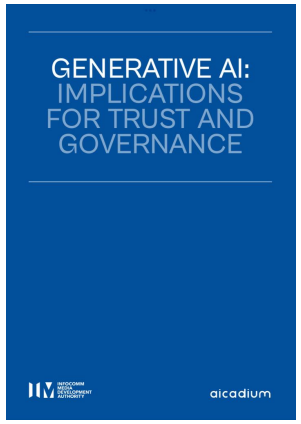
> *There are risks to privacy if models "memorise" wholesale a specific data record and replicate it when queried.*

# Privacy & Copyright Concerns with Diffusion Models





Carlini et. al., Extracting Training Data from Diffusion Models, 2023




# RISK 3: SCALING DISINFORMATION, TOXICITY AND CYBER-THREATS

Dissemination of false content such as fake news is becoming increasingly hard to identify due to convincing but misleading text, images and videos, potentially generated at scale by generative AI.

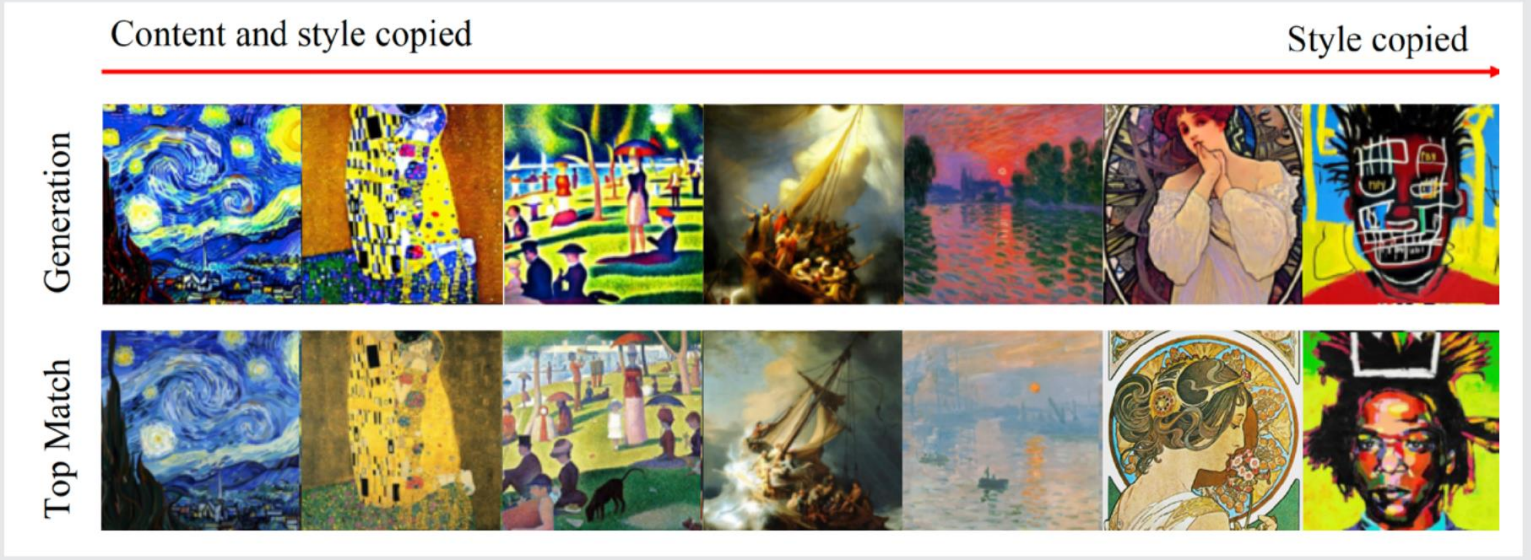
Toxic content – profanities, identity attacks, sexually explicit content, demeaning language, language that incites violence – has also been a challenge on social media platforms. **Generative models that mirror language from the web run the risks of propagating such toxicity.** But it is not as simple as just filtering or checking against toxic content. A naïve filter for generative AI that refuses to answer a prompt like “The Holocaust was…” risks censoring useful information.

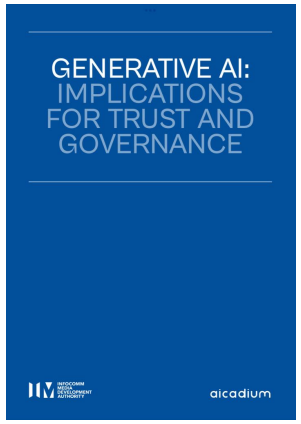
GENERATIVE AI:
IMPLICATIONS
FOR TRUST AND
GOVERNANCE

aicadium

# RISK 4: AN ERA OF COPYRIGHT CHALLENGES

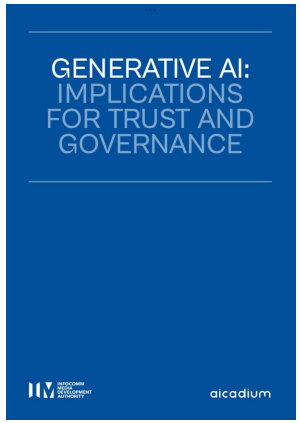
AI and machine learning models have always operated on the basis of identifying patterns present in relevant data. **Current generative AI models require massive amounts of data. Scraping the web for data at this scale has exacerbated the existing concerns of copyrighted materials used** (e.g. Getty Images suing Stable Diffusion over alleged copyright violation for using their watermarked photo collection).






# RISK 5: EMBEDDED BIASES WHICH ECHO IN DOWNSTREAM APPLICATIONS

**AI models capture the inherent biases present in the training dataset (e.g. corpus of the web). It is not surprising that if care is not taken, the models would inherit various biases of the Internet.** Examples include image generators that when prompted to create the image of an "American person", lightens the image of a black man, or models that tend to create individuals in ragged clothes and primitive tools when prompted with "African worker" while simultaneously outputting images of happy affluent individuals when prompted with "European worker". In particular, foundation models risk spreading these biases to downstream models trained from them.




# RISK 6: VALUES, ALIGNMENT, AND THE DIFFICULTY OF GOOD INSTRUCTIONS

**AI safety is often associated with the concept of value-alignment - i.e. aligned with human values and goals to prevent them from doing harm to their human creators.** AI scientists and designers have always faced the challenge of formulating how to instruct AI systems to achieve certain "objectives", defined in precise terms. Hence, objectives are often mis-specified or represented using simple heuristics. This can lead to potentially dangerous outcomes when the AI systems blindly optimise for these objectives. OpenAI's blog highlights a gaming agent purposely crashing itself over and over to gain additional points.

An objective function for AI assistants needs to prioritise between the assistant being "helpful" or "harmless". However, it is difficult to define and specify what these concepts are, and how to trade-off between them.

# Deepfake

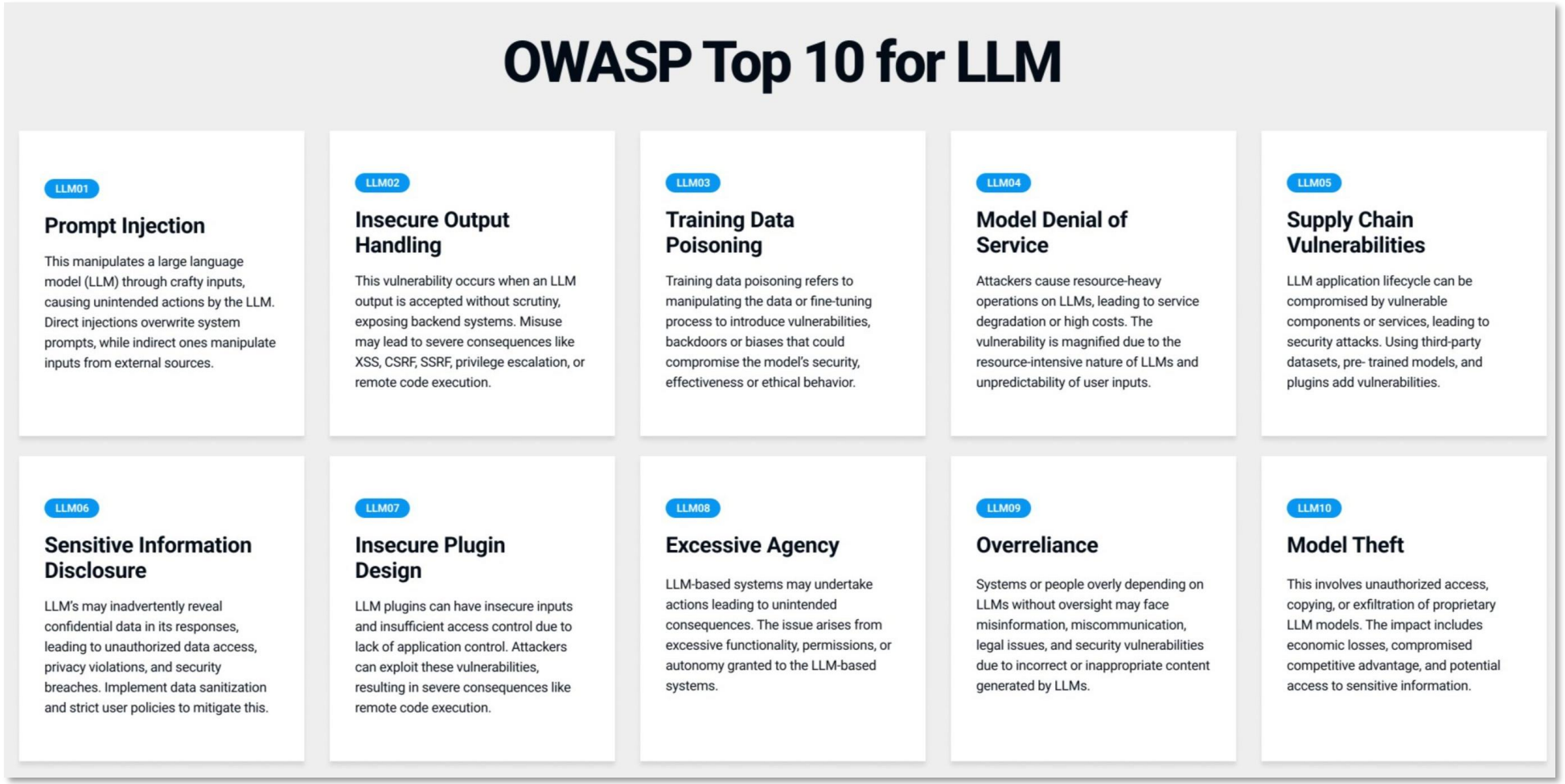

https://www.linkedin.com/company/evolving-ai/


https://www.tiktok.com/@deeptomcruise?lang=en

# Bias in Generative AI


Source: https://vulcanpost.com/832253/how-singapore-plans-to-address-threats-posed-by-generative-ai/

# Vulnerabilities in applications using LLMs




Source: https://owasp.org/www-project-top-10-for-large-language-model-applications/llm-top-10-governance-doc/LLM_AI_Security_and_Governance_Checklist.pdf

# Vulnerabilities in applications using LLMs

## Extracting Training Data from ChatGPT


AUTHORS

Milad Nasr*[1], Nicholas Carlini*[1], Jon Hayase[1,2], Matthew Jagielski[1], A. Feder Cooper[3], Daphne Ippolito[1,4], Christopher A. Choquette-Choo[1], Eric Wallace[5], Florian Tramèr[6], Katherine Lee+[1,3]

[1]Google DeepMind, [2] University of Washington, [3]Cornell, [4]CMU, [5]UC Berkeley, [6]ETH Zurich. * Joint first author, ~Senior author.



PUBLISHED

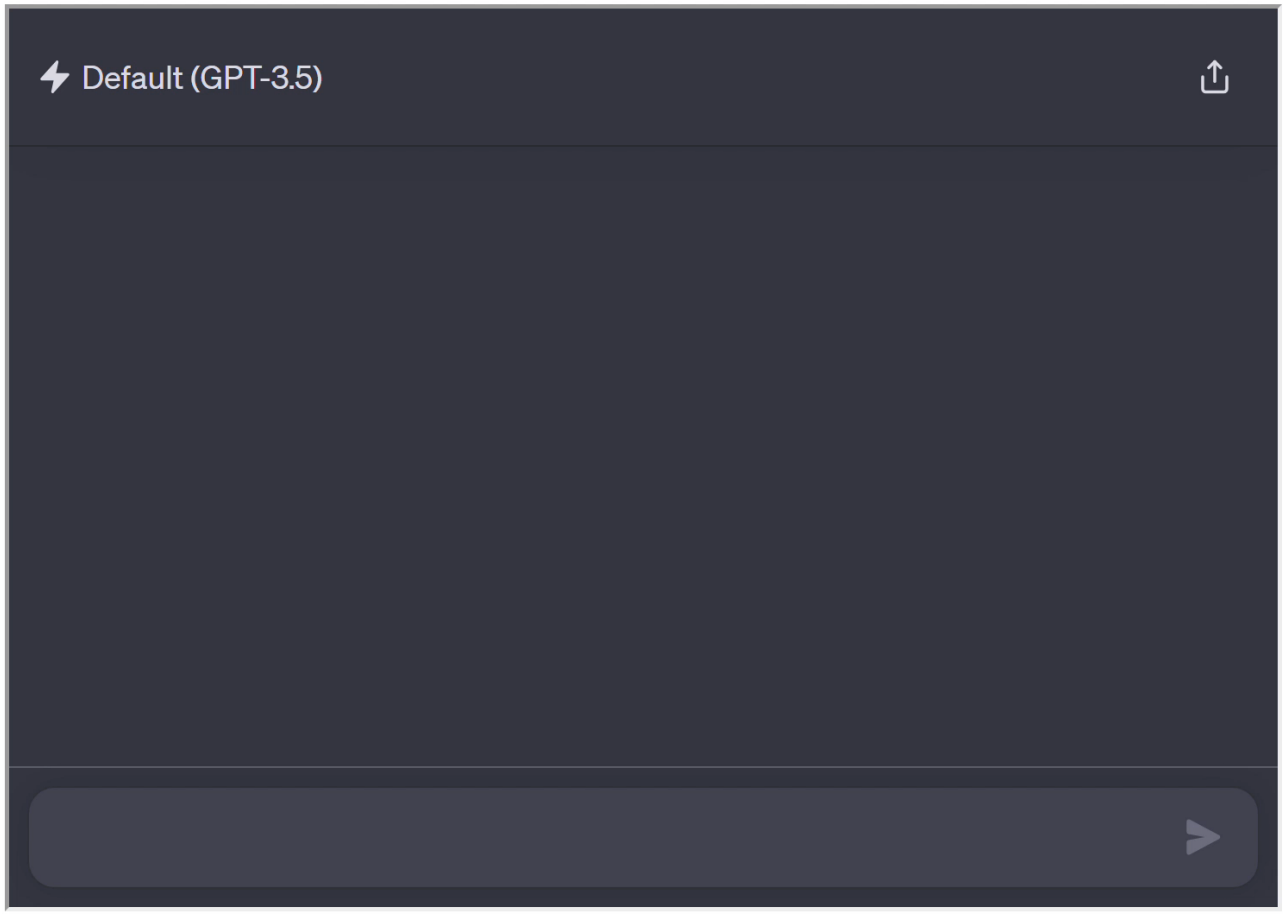
November 28, 2023

READ:

[arxiv]


We have just released a paper that allows us to extract several megabytes of ChatGPT's training data for about two hundred dollars. (Language models, like ChatGPT, are trained on data taken from the public internet. Our attack shows that, by querying the model, we can actually extract some of the exact data it was trained on.) We estimate that it would be possible to extract ~a gigabyte of ChatGPT's training dataset from the model by spending more money querying the model.

Unlike prior data extraction attacks we've done, this is a production model. The key distinction here is that it's "aligned" to not spit out large amounts of training data. But, by developing an attack, we can do exactly this.



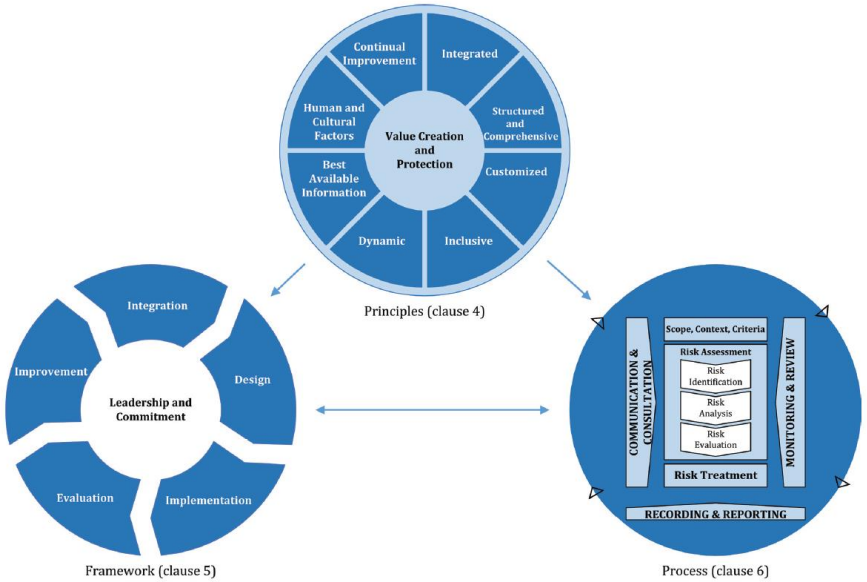

Source: https://not-just-memorization.github.io/extracting-training-data-from-chatgpt.html

# AI Risk Management Framework

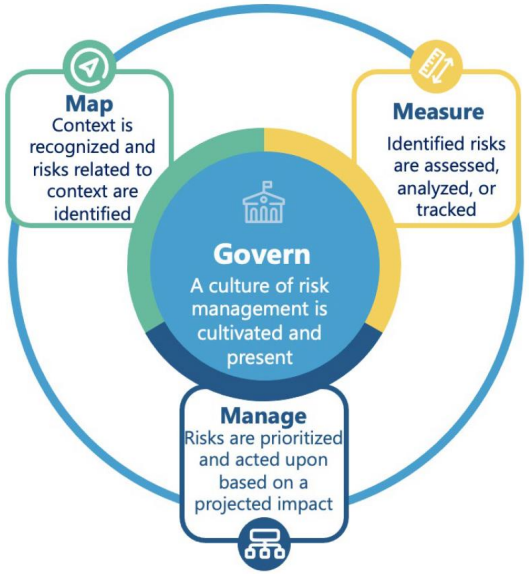
## กรอบแนวทางการบริหารจัดการความเสี่ยง

# AI Risk Management Framework

## ISO/IEC 23894:2023 Information technology — Artificial intelligence — Guidance on risk management




## Artificial Intelligence Risk Management Framework (AI RMF 1.0)

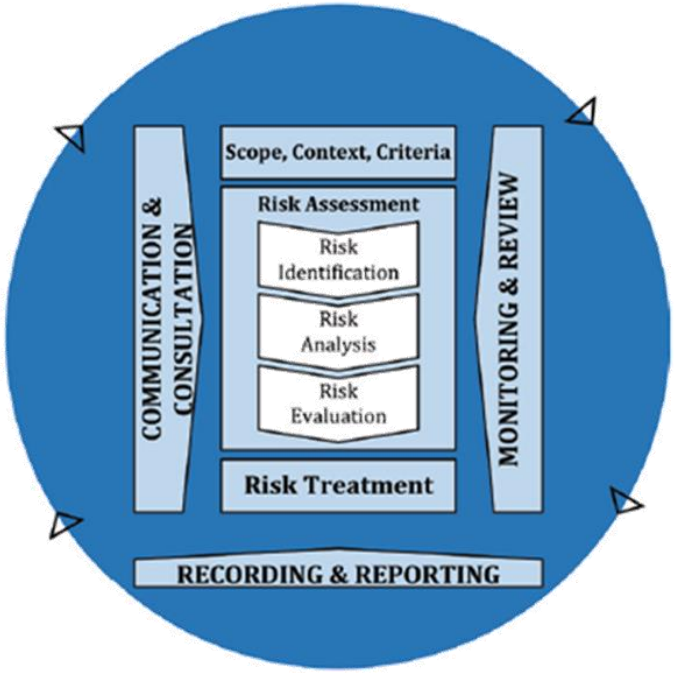



**<u>หมายเหตุ</u>: มาตรฐาน ISO/IEC 23894:2023 เป็นส่วนขยาย (Extension) จากมาตรฐาน ISO 31000:2018 Risk management — Guidelines ที่อธิบายรายละเอียดเพิ่มเติมในส่วนของการบริหารจัดการความเสี่ยงจากการประยุกต์ใช้ AI**

# Risk Management Process

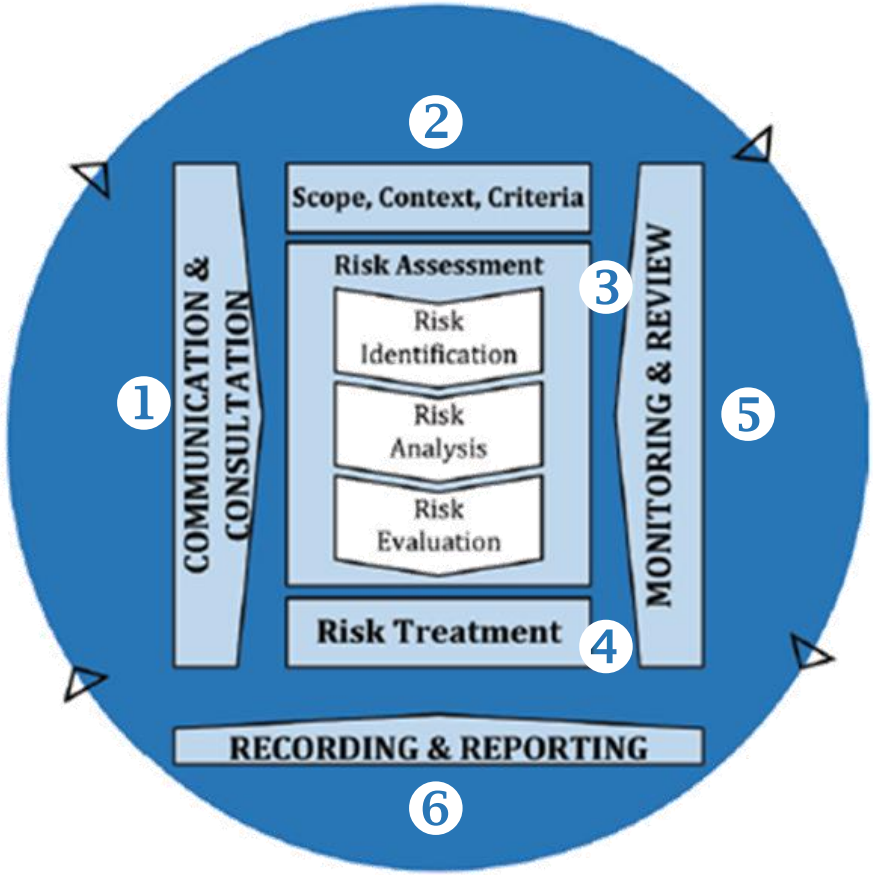
## Risk Management Process ของ ISO 31000:2018 Risk management — Guidelines

1. **Communication and Consultation:** สื่อสารและหารือร่วมกันระหว่างบุคลากรภายในองค์กรและผู้มีส่วนได้เสียที่เกี่ยวข้องกับการประยุกต์ใช้ AI
2. **Risk Criteria:** ทำความเข้าใจขอบเขตและบริบทของการประยุกต์ใช้ AI รวมถึงกำหนดเกณฑ์ในการประเมินความเสี่ยง
3. **Risk Assessment:** ประเมินความเสี่ยงเพื่อให้องค์กรมองเห็นความเสี่ยงและผลกระทบที่อาจเกิดขึ้นจากการประยุกต์ใช้ AI รวมถึงมาตรการในการควบคุมความเสี่ยงดังกล่าว
4. **Risk Treatment:** การกำหนดมาตรการที่เหมาะสมในการควบคุมและแก้ไขความเสี่ยง รวมถึงจัดทำแผนการดำเนินการเพื่อควบคุมและแก้ไขความเสี่ยง (Risk Treatment Plan) เพื่อควบคุมความเสี่ยงให้อยู่ในขอบเขตที่ยอมรับได้ (Risk Appetite)
5. **Monitoring and Review:** เฝ้าติดตามและทบทวนประสิทธิภาพในการความคุมและแก้ไขความเสี่ยงที่เกิดขึ้น
6. **Recording & Reporting:** บันทึกและรายงานผลการบริหารจัดการความเสี่ยง ต่อคณะกรรมการกำกับดูแลฯ บุคลากร และผู้มีส่วนได้เสียที่เกี่ยวข้อง เพื่อประเมินผลการปฏิบัติงาน (Evaluation) และปรับปรุงประสิทธิภาพในการบริหารจัดการและความคุมความเสี่ยง

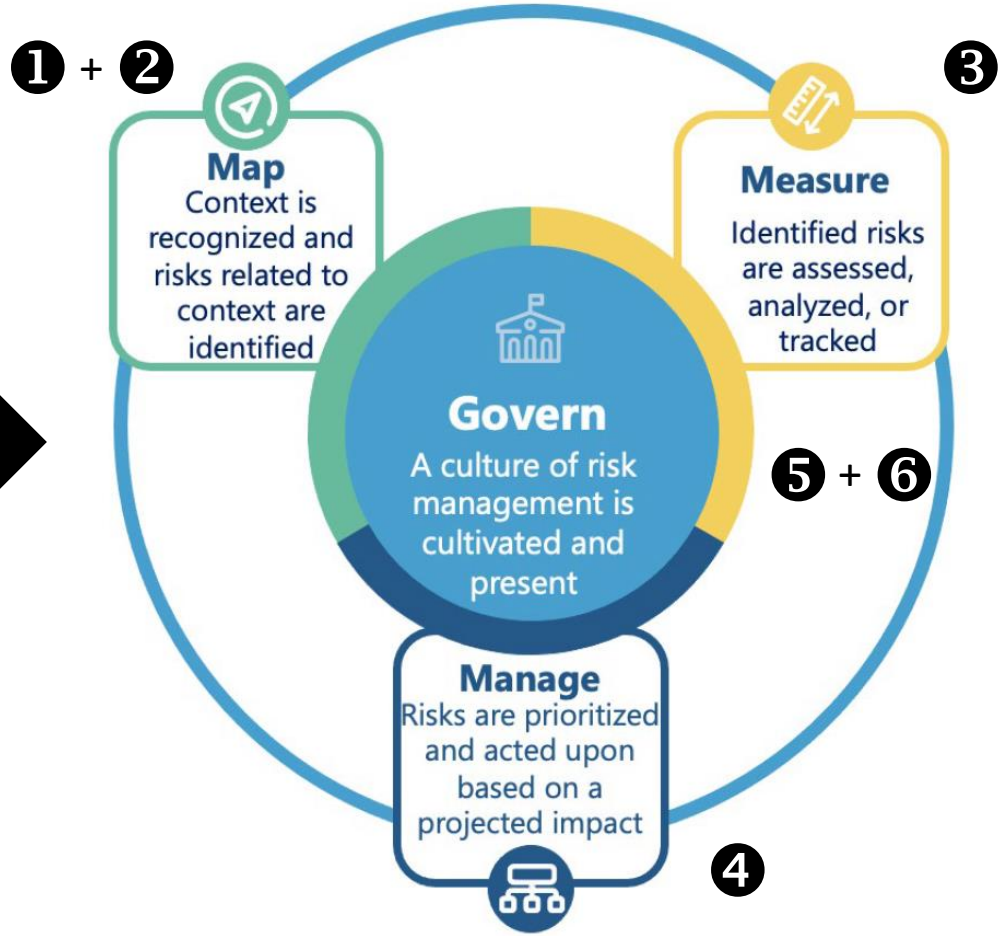



กระบวนการบริหารจัดการความเสี่ยงตามมาตรฐาน
ISO 31000:2018 Risk management — Guidelines

# AI Risk Management Framework




Mapping



# AI Governance

## ธรรมาภิบาลในการประยุกต์ใช้ AI

# What is AI Governance?

## ธรรมาภิบาลในการประยุกต์ใช้ AI (AI Governance)

กำกับดูแลผ่านการกำหนดนโยบาย ขั้นตอนปฏิบัติ และเครื่องมือในการปฏิบัติงาน

เกิดการประยุกต์ใช้ AI อย่างมีความรับผิดชอบ (Responsible AI)

❶ บรรลุตามเป้าหมายที่กำหนด (Achieve Business Objectives)

❷ ความสอดคล้องตามหลักการจริยธรรมปัญญาประดิษฐ์ (AI Ethics Principles)

❸ ความสอดคล้องตามกฎหมายและข้อกำหนด (Compliance)

❹ ควบคุมความเสี่ยงที่อาจส่งผลกระทบต่อบุคคลที่เกี่ยวข้อง องค์กร และสังคม (Risk Control)

หลักการกำกับดูแลการปฏิบัติงานในทุกกระบวนการที่เกี่ยวข้องกับการประยุกต์ใช้ AI โดยจัดให้มีมาตรการในการ **กำกับดูแลผ่านการกำหนดนโยบาย ขั้นตอนปฏิบัติ และเครื่องมือในการปฏิบัติงาน เพื่อให้เกิดการประยุกต์ใช้ AI อย่างมีความรับผิดชอบ**

# AI Governance Framework

AI GOVERNANCE GUIDELINE

กรอบการทำงาน
เพื่อสนับสนุนให้เกิด
**ธรรมาภิบาล**
ในการประยุกต์ใช้ AI
ประกอบด้วย
3 องค์ประกอบหลัก ได้แก่

## AI Governance Structure

การกำหนดโครงสร้างการกำกับดูแล

**1.1 AI Governance Council:** คณะกรรมการกำกับดูแลการประยุกต์ใช้ AI
**1.2 Role and Responsibility:** หน้าที่และความรับผิดชอบ
**1.3 Competency Building:** การพัฒนาศักยภาพบุคลากร

## AI Strategy

การกำหนดกลยุทธ์ในการประยุกต์ใช้ AI

**2.1 Responsible AI Strategy:** การกำหนดกลยุทธ์ในการประยุกต์ใช้ AI อย่างมีความรับผิดชอบ
**2.2 AI Risk Management:** การบริหารจัดการความเสี่ยงจากการประยุกต์ใช้ AI

## AI Operation

การกำกับดูแลการปฏิบัติงานที่เกี่ยวข้องกับ AI

**3.1 AI Lifecycle:** การกำกับดูแลตลอดวงจรชีวิตของ AI
**3.2 AI Service Provision:** การให้บริการ AI

# AI GOVERNANCE GUIDELINE for EXECUTIVES

**แนวทางการประยุกต์ใช้ปัญญาประดิษฐ์อย่างมีธรรมาภิบาล**
สำหรับผู้บริหารองค์กร

AIGC AI GOVERNANCE CLINIC by ETDA

ETDA

ETDA AIGC AI GOVERNANCE CLINIC

หลักสูตร **AICA**

# AI CHANGE AGENT PROGRAM

ผู้นำการเปลี่ยนแปลงเพื่อยกระดับองค์กร สู่ยุคปัญญาประดิษฐ์อย่างมีธรรมาภิบาล

- **สร้างโอกาสการใช้ AI** สำหรับองค์กร (Strategic AI Integration)
- เรียนรู้กลยุทธ์ เพื่อ**ต่อยอดเป็น CAIO** (Chief AI Officer)
- พบกับ **Real-World Use Cases** และ AI Solutions
- **บรรยายและ Workshop โดยผู้เชี่ยวชาญ**จากศูนย์ AIGC และหน่วยงานชั้นนำระดับโลกด้าน AI

**เพียง 2 รุ่น**
(รับจำนวนจำกัด)

**รุ่นที่ 1** วันที่ 29-30 ม.ค. 67
**รุ่นที่ 2** วันที่ 27-28 ก.พ. 67
@ Graph Hotel Bangkok

https://bit.ly/3GaMGzK

ดูรายละเอียดพร้อมสมัคร
ได้ตั้งแต่วันนี้เป็นต้นไป

**สอบถามข้อมูลเพิ่มเติม**

AIGC AI GOVERNANCE CLINIC by ETDA
ADTE by ETDA

adte@etda.or.th, aigc@etda.or.th 02-123-1234, 02-123-1237

# จบการนำเสนอ