September 2025 Issue

Taiwan and the Global AI Report-

Trust at Risk in AI Security

Author:
WANG Shih-Han

Editors:
Dr. CHEN Yen-Kuang

Dr. TUNG Chen-Yuan

Artificial intelligence (AI) has become an indispensable part of modern life, powering systems from conversational agents to autonomous vehicles and complex financial infrastructures. While these innovations offer tremendous opportunities, they also introduce complex risks that extend beyond technical vulnerabilities, affecting national security, economic stability, and public safety and trust.

The Global Alliance for Taiwan Technology Diplomacy (GATTD) seeks to facilitate cooperation, helping governments, industry, and academia confront these challenges together. Our mission is to strengthen global and regional security through research, partnerships, talent development, and commercialization, while fostering economic growth through collaboration between Taiwan and other technology-driven economies.In partnership with, and under the leadership of, the Taipei Representative Office in Singapore, we aim to share insights widely and connect audiences across Taiwan, Singapore, and beyond.

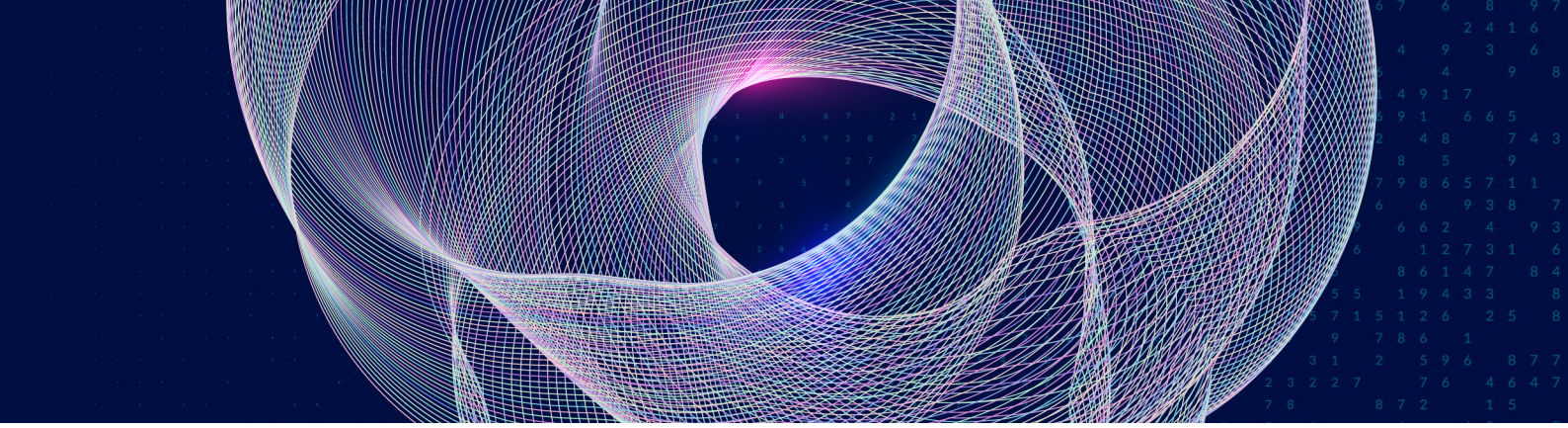
This inaugural report represents our first systematic effort, with an initial focus on AI security and safety. We convened leading Taiwanese AI experts to identify risks, outline safety protocols, and promote responsible practices. Our goal is to provide practical guidance for anticipating, mitigating, and managing these emerging AI risks effectively.

We welcome your feedback and comments, which will help us improve and expand future reports in this series.

Dr. CHEN Yen-Kuang
CEO, Global Alliance for Taiwan Technology and Diplomacy (GATTD)

Dr. TUNG Chen-Yuan
Representative, Taipei Representative Office in Singapore

For feedback purposes, please email Dr Chen at:
ykchen@gmail.com



Executive Summary

Artificial intelligence (AI) has rapidly evolved from a technological novelty into an indispensable component of modern life. Today, it powers systems ranging from conversational agents to autonomous vehicles and complex financial infrastructure.

This widespread adoption fuels extraordinary innovation but also simultaneously exposes society to a complex array of emerging risks. These risks extend far beyond technical vulnerabilities, affecting national security, economic stability, and the foundation of public safety and trust.

This report systematically introduces the main categories of AI security risks: privacy breaches, data manipulation, model exploitation (including adversarial attacks and prompt injection), malicious misuse, and unpredictable outputs, often referred to as “hallucinations.”

The urgency of addressing these risks is clear: trust in AI systems depends on their security and reliability. By highlighting tangible, real-world consequences, the report shows why the future of AI must prioritize proactive security, robust trust frameworks, and ethical safeguards. Collaborative risk mitigation across all sectors is essential to building a secure and trustworthy AI landscape.

Contents

1. Introduction: AI's Dual Nature – Innovation and Emerging Risks

The integration of AI into daily life marks a transformative era. At its core, AI refers to computer systems capable of performing tasks that typically require human intelligence, such as learning, problem-solving, and language understanding.

From large language models and AI-driven customer service to autonomous vehicles and advanced financial tools, AI is now deeply embedded in global workflows. These technologies demonstrate immense potential to foster innovation and efficiency.

However, this rapid advancement also introduces a new spectrum of risks. Traditional cybersecurity focuses on protecting networks and systems from digital attacks. AI security, in contrast, focuses on system vulnerabilities within the models themselves and in the data they rely on.

AI vulnerabilities are weaknesses in model design, data, or deployment that can be exploited to compromise function or safety. This report specifically examines the security and integrity of AI systems, showing how intrinsic vulnerabilities can be compromised and lead to unintended harm.

These concerns are deeply tied to national security, economic competitiveness, and public trust. The paradox is clear: the more deeply AI is woven into society, the greater the potential consequences of its failure. Understanding these risks is therefore a strategic priority for policymakers, institutions, and the public.

2. Overview of AI Risks: Understanding the Landscape

AI security is a complex, evolving discipline that affects the lifecycle of its digital infrastructure. To organize this discussion, the report classifies risks into five core categories. Each represents a distinct way AI systems can be undermined, they are deeply interconnected and collectively shape the broader threat landscape.

Privacy

The first category is **Privacy and Security**, whereby data breaches threaten sensitive personal data and private user conversations. Such incidents undermine user confidence and raise concerns about how responsible AI-driven systems are at handling information.

Data Poisoning

The second is **Data Poisoning and Integrity**, in which corrupted inputs manipulate models' interpretation of the world. For example, even a simple manipulated stop sign on the coding end can lead to a misinterpretation of an autonomous vehicle when making decisions on the road.

Model Exploitation

The third category, **Model Exploitation and Input Manipulation**, highlights attacks that occur during real-world use. From prompt injections that bypass safeguards to imperceptible tweaks that trick vision models, these exploits reveal that the adaptability that makes AI powerful can also be a critical weakness.

Deepfake AI

Fourth, the **Malicious Use of AI** weaponizes technology for harmful purposes. Deepfake fraud, automated phishing, and large-scale misinformation show how AI can amplify the impact of malicious actors far beyond traditional means.

Biases and Ethical Misalignments

Finally, **Unintended Consequences and Hallucinations** stem not from attackers but from AI itself. Overconfident falsehoods, embedded biases, and ethical misalignments demonstrate that even well-intentioned AI can perpetuate stereotypes and human biases.Together, these five categories illustrate the modern AI threat landscape. Addressing them requires not just technical safeguards but also vigilance, ethical design, and collaborative action across governments, industries, and communities.

Table 1. Key AI Risk Categories and Examples

Privacy and Security

Handling, storage, and transmission of sensitive data

Example:
User conversation exposure incidents (2023) [1]

Data Poisoning and Integrity

Malicious manipulation of training data to corrupt model behavior

Example:
Poisoned stop sign fooling self-driving cars [2][10][11]

Model Exploitation and Input

Manipulation at inference time (prompt injection, adversarial examples)

Example:
Prompt injection in large language models (LLMs) [3], sticker attack on stop sign

Malicious Use of AI

Repurposing AI for harmful activities such as fraud or disinformation

Example:
Deepfake scam causing $25M loss [4]

Unintended Consequences and Hallucination

AI producing false, biased, or unpredictable outputs

Example:
Fabricated cases cited by lawyer using an LLM [5]

These risks are present at every stage of the AI lifecycle, from data collection to deployment and ongoing operation. Unlike static vulnerabilities in traditional software, AI risks evolve continuously and demand adaptive solutions, robust governance, and sustained oversight. This makes AI security not a one-time problem to solve but an ongoing discipline that must evolve alongside the technology itself.

A. Privacy and Security: Safeguarding Personal Data in AI Systems

Privacy

AI systems rely on vast volumes of sensitive personal and organizational data, making privacy and security a central concern. Incidents such as user data leaks in 2023, which exposed private conversations, demonstrate the risks inherent in data storage and transmission [1]. Beyond accidental leaks, more sophisticated threats such as **model inversion attacks** enable adversaries to reconstruct private training data simply by analyzing a model's outputs [12].

These risks are not limited to individuals. In some cases, corporate employees have inadvertently shared proprietary information—such as source code or meeting minutes—with publicly accessible AI models [1]. Such exposures underscore that privacy risks extend across personal, organizational, and national levels.

These events raise critical questions about user consent, data transparency, and the ethics of AI data use. If people and organizations cannot trust that their information is handled responsibly, adoption and confidence in AI technologies will inevitably suffer.

Mitigation strategies include **differential privacy**, which introduces statistical noise to datasets while preserving useful patterns, and **federated learning**, which enables distributed model training without centralizing sensitive data. Together, these approaches reduce the risks of exposure while maintaining the utility of AI.

B. Data Poisoning and Algorithm Attacks: Undermining AI Integrity

Data Poisoning

The reliability of AI systems depends on the integrity of their training data and algorithms. **Data poisoning** occurs when malicious inputs are inserted into datasets, subtly corrupting model behavior.

A striking example is provided by Eykholt et al. (2017), who showed that stickers placed on a stop sign could cause an autonomous vehicle's vision system to misclassify it as a speed limit sign [11][13].

While this may seem simple, the implications for safety are profound: corrupted inputs can endanger lives.

Other forms of poisoning include **label flipping**, where data labels are maliciously reassigned, and the injection of biased samples, which lead to discriminatory or unfair outputs. These evolving threats emphasize the need for ongoing data audits, validation pipelines, and continuous monitoring of deployed systems, because unlike attacks on deployed models, data poisoning corrupts the system's core understanding of the world before it even enters service, making the flaw difficult to detect later.

C. Model Exploitation and Input Manipulation: Prompt Injection and Adversarial Attacks

Model Exploitation

Beyond training data, modern AI models are also vulnerable during real-world operation. Two prominent classes of attacks, **prompt injection and adversarial manipulation**, demonstrate how easily AI behavior can be subverted.

Prompt injection primarily targets **large language models (LLMs, a type of AI built on deep neural networks that generate human-like text).** By embedding hidden instructions into user inputs, attackers can override safeguards to extract sensitive data. For example, a crafted prompt might instruct the system to ignore safety rules and spread harmful or confidential information [3].

Adversarial attacks, on the other hand, exploit imperceptible **perturbations—** tiny, deliberate changes to input data—to mislead AI systems. These are different from random noise, which is accidental; perturbations are carefully designed so that humans see little or no difference, but translate into an entirely different message to AI. Figure-1 [14] shows how an adversarial perturbation can transform an image of a panda into a misclassified "gibbon." Figures-2 and -3 [13] extend this to the physical world, where stickers on stop signs cause autonomous vehicles to misinterpret them. Figure-4 [13] reveals how such manipulations—nearly invisible to humans—produce high-confidence but incorrect classifications.

These attacks highlight a fundamental paradox: the very adaptability that makes AI powerful also creates avenues for exploitation. Addressing them requires robust defenses, adaptive monitoring, and resilience against both digital and physical manipulation.

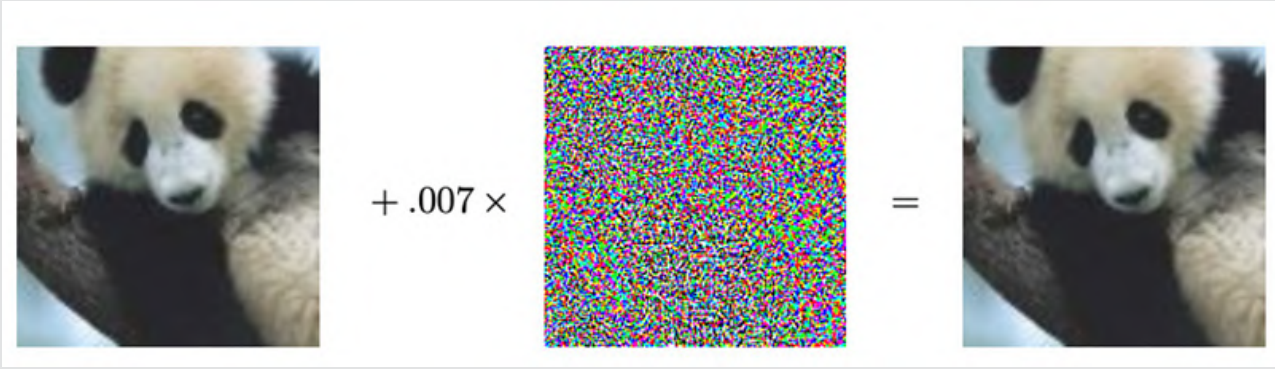


Figure-1: Adversarial Examples in the Digital World: When AI Sees What Humans Do Not [14]

Figure-1 illustrates how artificial intelligence systems can be manipulated through what are called adversarial examples. On the left, the AI correctly recognizes a panda. In the center is a pattern of carefully engineered adversarial perturbation that appears meaningless to the human eye. When this tiny, nearly invisible alteration is added to the original image (right), the AI is misled into classifying the panda as a "gibbon" with very high confidence while human observers see no meaningful differences.

This example highlights a key risk: AI systems can be entirely wrong. Small, hidden manipulations can exploit these weaknesses to signal incorrect information that goes hidden under the radar. For regulators and policymakers, it underscores the importance of ensuring that AI used in critical applications, such as security, healthcare, or transportation, has safeguards against such vulnerabilities.

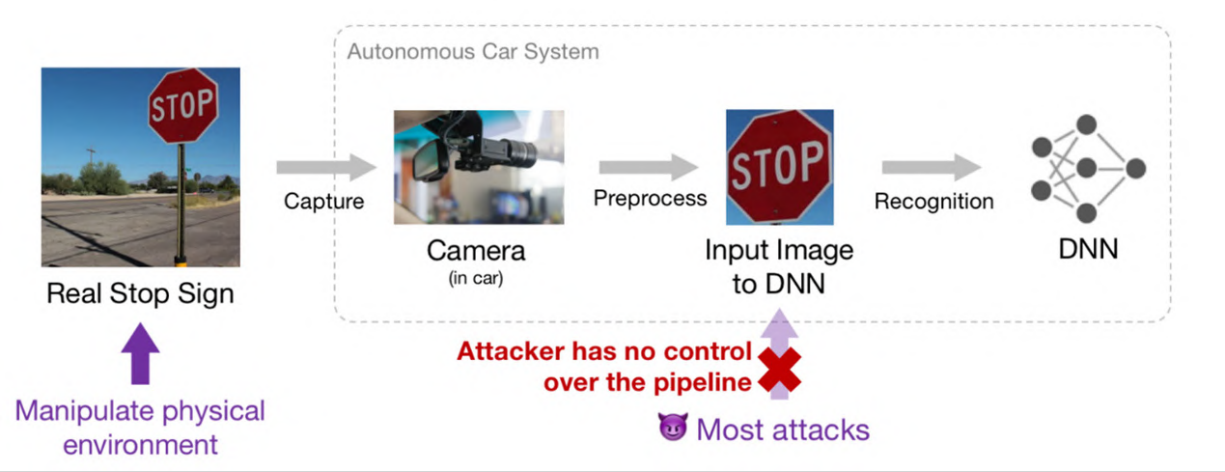


Figure-2: Adversarial Attack in the Real World: When Stop Signs Become an Attack Target. [13]

Unlike the previous example in Figure-1, which demonstrated adversarial manipulation in the digital world (adding adversarial perturbation to an image of a panda), Figure-2 shows how attacks can also occur in the physical world. On the left, a real stop sign is presented. An autonomous vehicle's camera captures the image, which is then processed by its deep neural network (DNN: a type of AI model inspired by the human brain, designed to recognize patterns such as objects, text, or faces) for recognition. Although attackers have no direct control over the car's internal pipeline, they can manipulate the physical stop sign itself, for example, by placing patterns on it shown as Figure-3.

Figure-3: Adversarial Manipulation Example in the Real World: Risks for Autonomous Vehicles [13]

Figure-3 shows how small physical changes, which show no meaningful difference to human observers, can mislead an autonomous vehicle's vision system across video frames. While humans still recognize the sign, the AI shifts from correctly identifying it as a stop sign to misclassifying it as a "sports ball." This highlights that adversarial attacks are not only digital but can also occur in the real world and unfold dynamically in driving conditions, leading to dangerous errors.

Figure-4: Adversarial Stop Signs in the Real-World: What Manipulation Looks Like [13]

Following Figure-3, which showed how an autonomous vehicle can misinterpret a stop sign across video frames, Figure-4 illustrates what those manipulations actually look like. To human eyes, each image still appears to be a normal stop sign. But the faint, texture-like patterns trick the AI into misclassifying them—as a person (left), a sports ball (middle), or not a stop sign at all (right).

Adversarial manipulation is not abstract but can be physically realized with small, strange-looking patterns. These changes are usually ignored by human drivers yet can cause AI systems to make high-confidence but dangerous errors in real-world driving.

D. Malicious Use of AI: Exploiting Technology for Harm

Deepfake AI

While some risks stem from vulnerabilities, others arise from deliberate abuse of AI. Criminals and malicious actors are increasingly repurposing AI to amplify fraud, cybercrime, and disinformation.

Deepfake scams impersonating executives or family members have already caused multimillion-dollar financial losses [4]. **Voice impersonation** [6] enables convincing social engineering, while AI-generated phishing campaigns drastically increase both the scale and believability of attacks.

A 2024 review highlighted that phishing and compromised credentials remained the leading causes of data breaches, with the average U.S. breach cost reaching $9.36 million [7]. These examples show that AI's threats extend far beyond technical systems; they directly endanger individuals, businesses, and entire industries.

Mitigation therefore requires not only technological safeguards but also **digital literacy** and **critical thinking** to recognize and resist these emerging threats.

Biases and Ethical Misalignments

E. Unintended Consequences and Hallucination: The Unpredictable Side of AI

Not all risks come from malicious intent. Many are rooted in AI's **inherent unpredictability** and the biases in its training data.

Hallucination occurs when models confidently generate plausible but false information, such as the 2023 incident where a lawyer submitted fabricated cases produced by an LLM [5]. **Bias and discrimination** can also emerge, leading to unfair hiring practices or wrongful arrests [8][9].

Even when technically correct, AI outputs may be **ethically misaligned**, offering misguided advice or violating human moral values. The complexity and scale of AI systems also raise concerns about **systemic risks**, where cascading failures could spread unpredictably across domains.

These outcomes underscore the necessity of human oversight, continual auditing, and "human-in-the-loop" approaches—especially in high-stakes domains such as healthcare, law, and transportation.

3. Real-World Impact: Case Studies in AI Security

The risks outlined are not theoretical. Real-world incidents have already demonstrated how vulnerabilities in AI systems can translate into significant harm.

Table 2 highlights case studies ranging from user data leaks and adversarial stop sign attacks to multimillion-dollar deepfake scams and biased hiring algorithms. These events illustrate the **breadth of AI risks,** from privacy and financial harm to erosion of trust and systemic failure.

Collectively, these cases emphasize that AI security challenges are active, pressing, and growing. They call for proactive defenses, regulatory frameworks, and cross-sector collaboration to ensure AI technologies remain reliable and safe.

Table 2. Real-World AI Security Incidents and Their Impact

Incident/Case Study	Risk Category	Brief Description of Impact
User Data Leak Incidents (2023)	Privacy and Security	Exposure of sensitive user conversations or data [1]
Inadvertent Corporate Data Leak via Public AI Models	Privacy and Security	Accidental leak of proprietary information, such as source code or meeting minutes, to publicly accessible AI systems [1]
Adversarial Attack on Stop Sign (Eykholt et al., 2017)	Model Exploitation and Input Manipulation	Stickers cause self-driving car vision systems to misclassify signs [11]
Prompt Injection Attacks on LLMs	Model Exploitation and Input Manipulation	Malicious prompts override AI instructions, reveal data, or bypass safety features [3]
Deepfake Scam (Multinational Firm, $25M loss)	Malicious Use of AI	A multinational firm was defrauded after an employee participated in a video call featuring deepfake versions of the CFO and other staff, resulting in a $25 million transfer [4]
AI Voice Scam (Florida Woman, $15K loss)	Malicious Use of AI	Loss via AI-cloned voice, family impersonation [6]. A Florida woman lost $15,000 when criminals used AI voice cloning to impersonate her daughter in distress
Lawyer Citing Fabricated Cases from an LLM (2023)	Unintended Consequences / Hallucination	Misinformation, leading to legal and professional harm [5]
Chatbot Providing Incorrect Information	Unintended Consequences / Hallucination	AI chatbot providing erroneous advice, leading to financial loss or legal disputes for users [5]
AI Bias in Hiring/Law Enforcement	Unintended Consequences / Hallucination	Biased outcomes, wrongful arrests, or unfair discrimination due to algorithmic bias [8][9]
AI Chatbot Suggesting Harmful Advice	Unintended Consequences / Hallucination	Dangerous or unethical chatbot responses to vulnerable users [8]

Conclusion

Artificial intelligence is no longer a distant possibility; it is now embedded in systems that shape economies, societies, and daily lives. This report has outlined the key categories of AI risks: threats to privacy, data poisoning, adversarial input manipulation, malicious misuse, and unintended consequences such as hallucinations and bias. Each of these risks is real and evidenced by tangible examples, from multimillion-dollar deepfake scams to manipulated traffic signs. Together, they underscore the vulnerabilities that must be addressed to safeguard trust in AI.

Ultimately, AI security is not about eliminating all risk but about managing it responsibly. With vigilance, cross-sector collaboration, and a commitment to ethical design, we can build AI systems that are not only powerful but also safe, transparent, and reliable. By anticipating vulnerabilities and learning from past incidents, society can harness AI's transformative potential while safeguarding against its potential pitfalls.

References

[1] ChatGPT Data Leaks and Security Incidents (2023-2025): https://wald.ai/blog/chatgpt-data-leaks-and-security-incidents-20232024-a-comprehensive-overview

[2] Data Poisoning Attacks: How AI Models Can Be Corrupted, Lumenova AI Blog: https://www.lumenova.ai/blog/data-poisoning-attacks/

[3] Prompt Injection: Can a Simple Prompt Hack Your LLM? - G2 Learning Hub: https://learn.g2.com/prompt-injection?hsLang=en

[4] CyberheistNews Vol 14 #07 Social Engineering Masterstroke: How Deepfake CFO Duped a Firm out of $25 Million: https://blog.knowbe4.com/cyberheist-news-vol-14-07-social-engineering-masterstroke-how-deepfake-cfo-duped-a-firm-out-of-25-million

[5] Lawyer in Huge Trouble After He Used ChatGPT in Court and It Totally Screwed Up: https://futurism.com/the-byte/lawyer-chatgpt-court

[6] Florida Woman Loses $15K to AI Voice Scam - Bitdefender: https://www.bitdefender.com/en-us/blog/hotforsecurity/florida-woman-loses-15k-to-ai-voice-scam-mimicking-daughter-in-distress

[7] The 2024 Year in Review: Cybersecurity, AI, and Privacy Developments - Hinckley Allen: https://www.hinckleyallen.com/publications/the-2024-year-in-review-cybersecurity-ai-and-privacy-developments/

[8] AI poses threats of discrimination and violations of civil liberties, ACLU: https://www.newsfromthestates.com/article/ai-poses-threats-discrimination-and-violations-civil-liberties-aclu-says

[9] AI is biased against speakers of African American English, study finds: https://news.uchicago.edu/story/ai-biased-against-speakers-african-american-english-study-finds

[10] BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain: https://arxiv.org/pdf/1708.06733

[11] Robust Physical-World Attacks on Deep Learning Visual Classification (Eykholt et al., 2017): https://arxiv.org/pdf/1707.08945

[12] Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures: https://rist.tech.cornell.edu/papers/mi-ccs.pdf

[13] ShapeShifter: Robust Physical Adversarial Attack on Faster R-CNN Object Detector: https://arxiv.org/pdf/1804.05810

[14] Explaining And Harnessing Adversarial Examples: https://arxiv.org/pdf/1412.6572

Appendix A: Key Technical Terms

- ***Artificial Intelligence (AI):*** The field of computer science focused on building systems that can perform tasks requiring human-like intelligence, such as learning from data, recognizing patterns, reasoning, understanding language, and making decisions. Today's AI spans from narrow systems like chatbots or fraud detectors to advanced models capable of generating text, images, or analyzing complex data.
- ***Deep Neural Network (DNN):*** A type of AI model inspired by the structure of the human brain, built from layers of interconnected "neurons." DNNs are the dominant architecture in modern AI, powering most cutting-edge applications, from computer vision and speech recognition to large language models (LLMs) and generative AI systems like ChatGPT or image generators. Their ability to learn complex patterns from massive datasets makes them the foundation of today's AI revolution.
- ***Large Language Model (LLM):*** A specialized type of AI built on DNN architectures, trained on vast amounts of text to understand and generate human-like language. LLMs (such as GPT models) power many generative AI systems, enabling applications like chatbots, automated writing, code generation, and translation.
- ***Prompt Injection:*** A method of tricking an AI system, especially LLMs, by embedding hidden instructions inside user inputs. Prompt injection can override safeguards, extract sensitive data, or force the model to behave in unintended ways.
- ***Noise:*** Random changes or errors in data that happen naturally, such as camera blur or background sounds. Noise is not intentional but can still make AI less accurate.
- ***Perturbation:*** A small, intentional change made to trick or test an AI system. For example, tiny edits to an image that humans don't notice but cause the AI to misidentify what it sees.
- ***Adversarial Perturbation:*** A special kind of perturbation crafted by an attacker to deliberately fool AI systems while remaining almost invisible to humans. Unlike ordinary noise, which is random, adversarial perturbations are precisely engineered so the AI produces confident but wrong predictions—for instance, misclassifying a stop sign as a speed-limit sign.

- ***Adversarial Example:*** An input that has been subtly altered with an adversarial perturbation—a small, deliberate change designed to fool AI systems. To humans the input looks unchanged, but the AI misclassifies it with high confidence. Unlike random noise, which happens by accident, adversarial perturbations are intentional and carefully engineered to exploit weaknesses in the model.
- ***Federated Learning:*** A training method that allows AI models to learn from decentralized data (spread across devices or servers) without transferring sensitive information to a central location.
- ***Differential Privacy:*** A statistical technique that protects individual data points by adding "noise" to datasets, ensuring privacy while preserving overall patterns.
- ***Model Inversion Attack:*** An attack where adversaries reconstruct sensitive training data (e.g., faces, medical records) from the outputs or confidence scores of a machine learning model.
- ***Hallucination:*** When an AI generates outputs that sound plausible but are actually false, biased, or misleading.

Please feel free to reach out to the Economic Division of the Taipei Representative Office in Singapore should you have any enquiries or are seeking partnership opportunities of investment or collaboration in the field of semiconductors and AI in Taiwan.

Email: singapore@sa.moea.gov.tw
Telephone: +65 6500-0128

Published: Taipei Representative Office in Singapore
Address: 460 Alexandra Road, #23-00 mTower, Singapore 119963
Email: sgp@mofa.gov.tw
Telephone: +65 6500-0100

Design: Serena, Fang Ching Liu