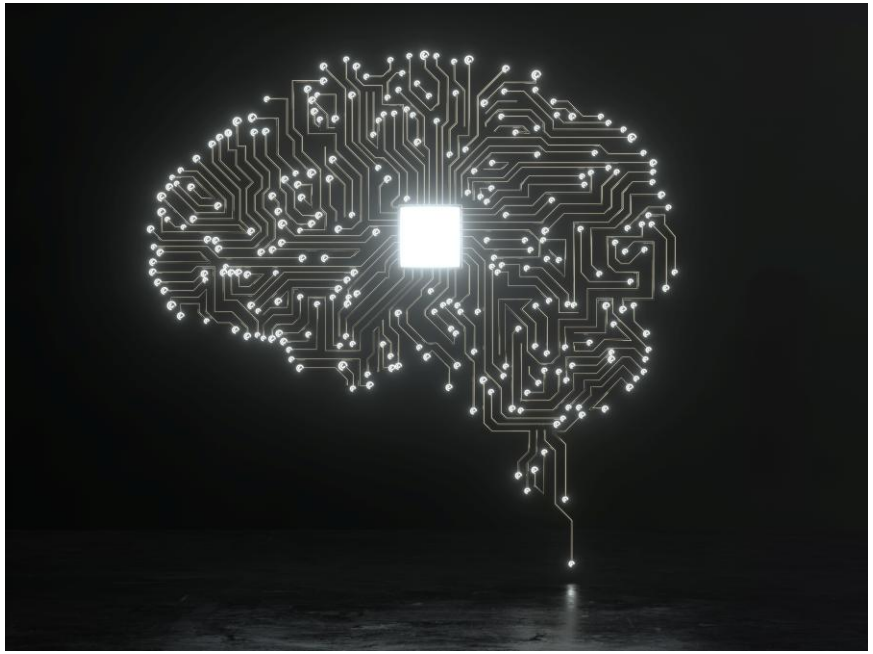
Agentic AI *from* a Cybersecurity Perspective

Presented by Pentastic Security Limited

Miranda Ma, Director, Business Development and Data Privacy
Wilson Fong, Senior Consultant, IT Security Awareness Training

Embracing AI in a Sea of Rapid Change

3.1Pro 3.6 V4 4.3 4.6 4.7 5.1 5.5



Source: *Microsoft 365 Stock Image*

Source: *AI-generated illustration (created using Microsoft Copilot)*

Agentic AI – A Paradigm Shift, Not Just Another Tool

What industry leaders are saying about Agentic AI

> *When we gathered at Next a year ago, we talked about how generative AI was transforming organizations around the world.* ***Today, that future is in-production—the Agentic Enterprise is real--and deployed at a scale the world has never before seen.***

Thomas Kurian
CEO, Google Cloud

Source: Thomas' keynote remarks for Google Cloud Next '26

> ***Future workforces* in enterprise will be *a combination of humans and digital humans.***

Jensen Huang
CEO, Nvidia

Source: https://fortune.com/2025/10/20/jensen-huang-nvidia-ai-future-workforce-digital-humans-hiring-onboarding-orientation/

THE FUTURE OF AGENTS

AI is about to completely change how you use computers

And upend the software industry.

By **Bill Gates** published on Friday, Nov 10, 2023

Work

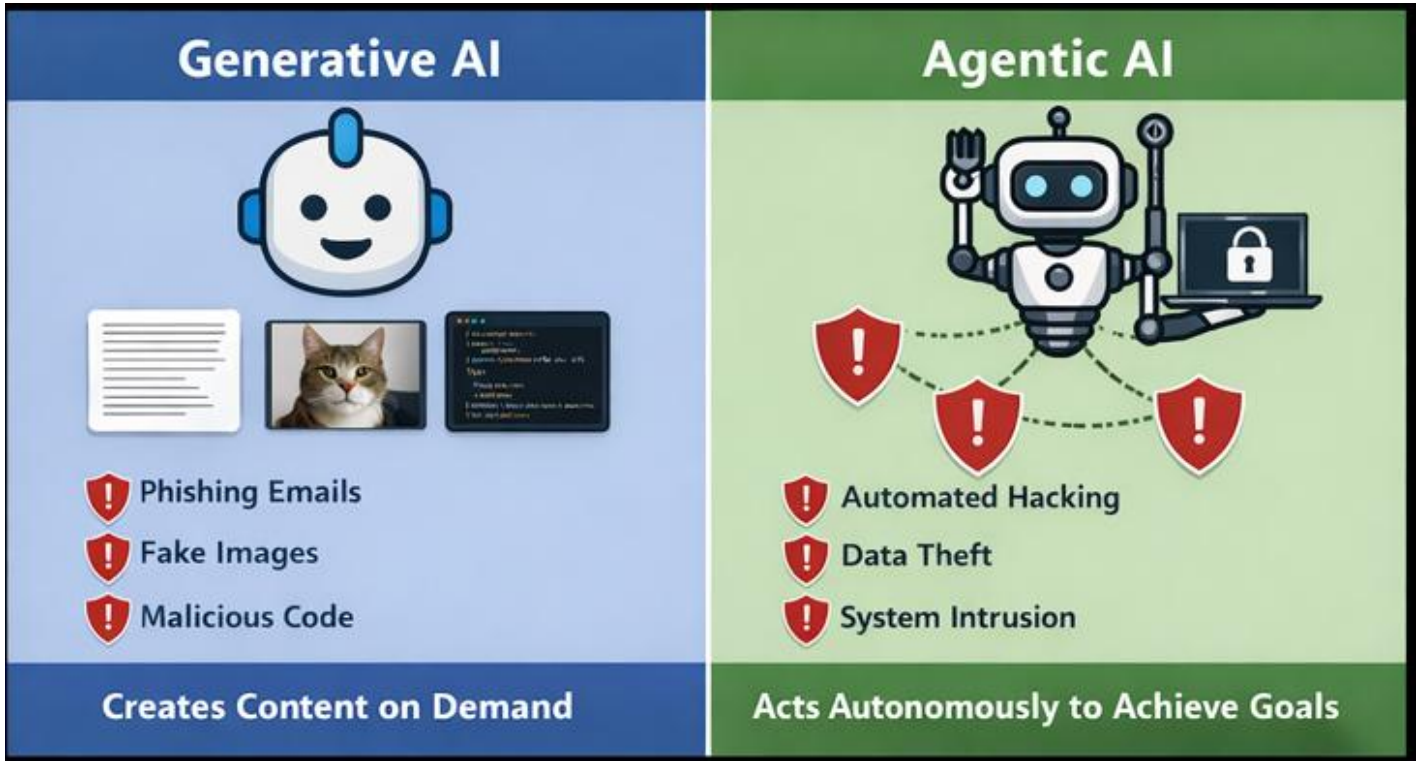
> ***Agents*** *are not only going to change how everyone interacts with computers. They're also going to upend the software industry,* ***bringing about the biggest revolution in computing*** *since we went from typing commands to tapping on icons.*

Bill Gates
Co-founder, Microsoft

Source: https://www.gatesnotes.com/ai-agents

From Content Creation to Autonomous Action

Generative AI creates new content, such as text, images, video, audio, or code, in response to user's prompts.



Source: *AI-generated illustration (created using Microsoft Copilot)*

Agentic AI autonomously plan, decide, and take actions to achieve a goal, with minimal human supervision.

When AI works against us, both Generative AI and Agentic AI are creating new cyber threats!

Illustration of an Agentic AI Execution Flow

Agentic AI Requires More Than an LLM — It Needs a Brain, Tools, and Memory

Set the Goal (by Human): Monitor Cybersecurity News

Set the Goal

Plan the Tasks

(1) Search everyday cybersecurity news automatically,

(2) Summarise incidents, repeats daily on schedule,

Plan the Tasks

Act Autonomously

(3) Searches, filters, summarises, emails

Act Autonomously

Deliver Results

(4) Send a daily email to your inbox at 9:30 am.

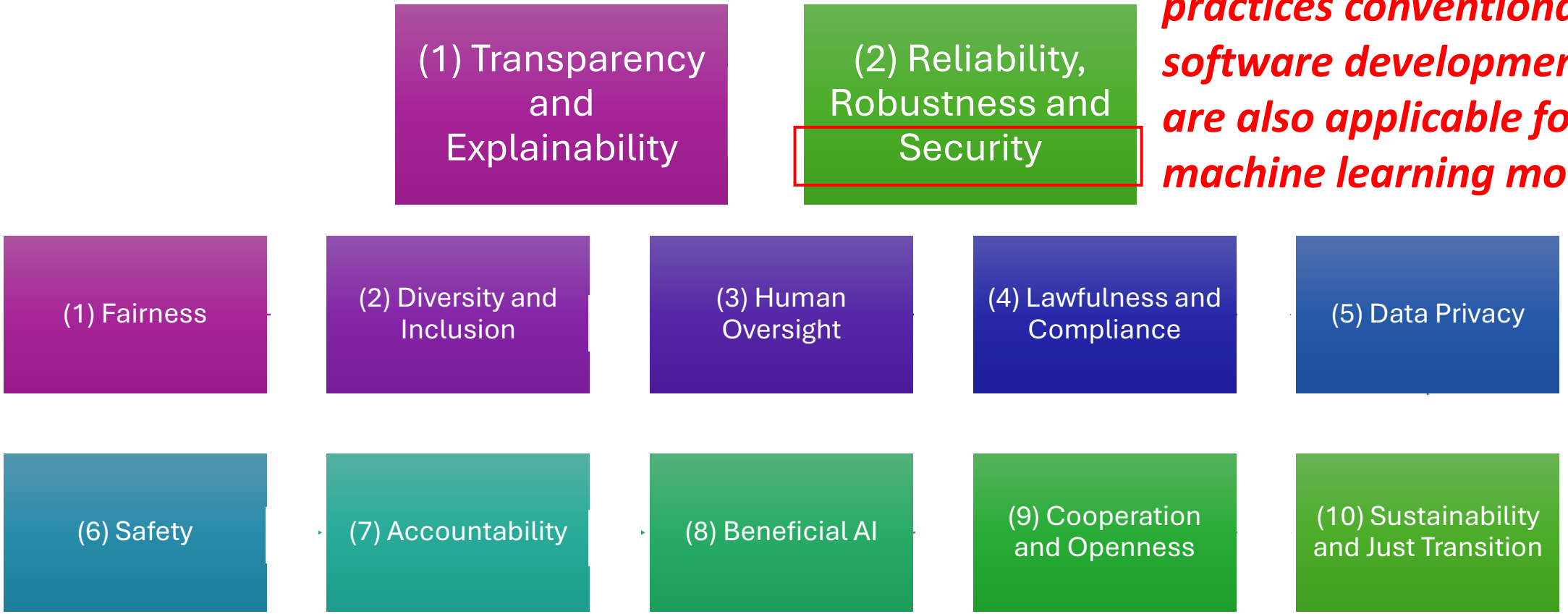
Deliver Results

Source: *AI-generated illustration (created using Microsoft Copilot)*

Security is Necessary but Not Sufficient for Ethical AI

Ethical AI Framework: *Two Performance Principles and 10 General Principles*

Many of existing security practices conventional to software development efforts are also applicable for AI and machine learning models.



Source: Ethical Artificial Intelligence Framework Version 2.0 dated December 2025, By Digital Policy Office

From Concept to Reality

What's next in today's agenda

Level One: AI Security Awareness
for AI Application/Service End User

Level Two: Security Testing
for Corporates Developing and/or Deploying AI Application/Service

Source: *AI-generated illustration (created using Microsoft Copilot)*

Level One

AI Security Awareness

for AI Application/Service End User

Poor Prompt Quality may lead to Unintended results

Challenge #1

Why this matters to users
If you instruct an agent incorrectly, vaguely, or carelessly, the AI may still execute actions—**at machine speed and scale**—with real operational impact.
“I let the AI do it” is **not** a defensible explanation during incidents, compliance reviews, or investigations.

Takeaway
Prompt ambiguity increases the risk of privilege misuse, misaligned actions, and operational errors.

Our recommendations for users:
Learn using prompts effectively!
Explicitly set limits on scope, data sources, and permissible actions.

Source: *AI-generated illustration (created using Microsoft Copilot)*

Over-Sharing Information may create Long-Lived Risk

Challenge #2

Why this matters to users
Sensitive data (e.g. personal data, credentials, internal documents) may be reused, inferred, or acted upon later—well beyond the original interaction

Takeaway
Assume anything you share with an agent may affect future decisions. It may cause personal data leakage.

Our recommendations for users:
Review documents for sensitive elements before inputting into AI tools and Use Anonymization technique;
Extract only necessary excerpts;
Swap names or redact;
Only use DLP (Data Loss Prevention) as last line of defence

Source: *AI-generated illustration (created using Microsoft Copilot)*

Over-Trust in AI-authority may amplify manipulation risks

Challenge #3

Why this matters to users
Users may treat AI outputs as authoritative, especially when phrased confidently.
Agentic workflows may propagate incorrect outputs into downstream systems.

Takeaway
AI confidence does not equal correctness.
“AI-generated content may be incorrect”

Our recommendations for users:
Human verification (human-in-the-loop) is mandatory especially in the high risk AI systems.

Source: *AI-generated illustration (created using Microsoft Copilot)*

Non-Approved Agentic AI Platforms lack Security Visibility and Control

Challenge #4

Why this matters to users

Non-approved agentic AI platforms operate outside standard monitoring, logging, and access controls, creates blind spots that attackers can exploit and prevents effective incident detection and response.

Source: *AI-generated illustration (created using Microsoft Copilot)*

Takeaway

If the platform or agent is not officially approved, it must be treated as untrusted — regardless of popularity, do not use it for work tasks.

Our recommendations for users:

For work, follow your corporate's AI policy and procedures
For personal use, using AI agent in an isolated environment.
Treat third-party tools/skills as untrusted and check before use.

Level One Takeaway

- Agentic AI does not replace basic security hygiene. Existing security best practices—such as strong password management, email security, secure use of websites and mobile apps, and safe use of Wi-Fi—remain fully applicable.

- Before delegating tasks to Agentic AI: Think before assigning actions, limit permissions and scope, review outputs and actions, and own the outcomes.

- Agentic AI can operate autonomously; however, responsibility for secure and ethical use remains with the user and the organisation.

Source: *AI-generated illustration (created using Microsoft Copilot)*

Level Two

Security Testing

for Corporates Developing and/or Deploying AI Application/Service

OWASP Top 10 for Agentic Applications

The most critical security risks associated with Agentic AI Applications

OWASP (Open Worldwide Application Security Project) is a global, non-profit community that publishes **industry-accepted security risk frameworks**.

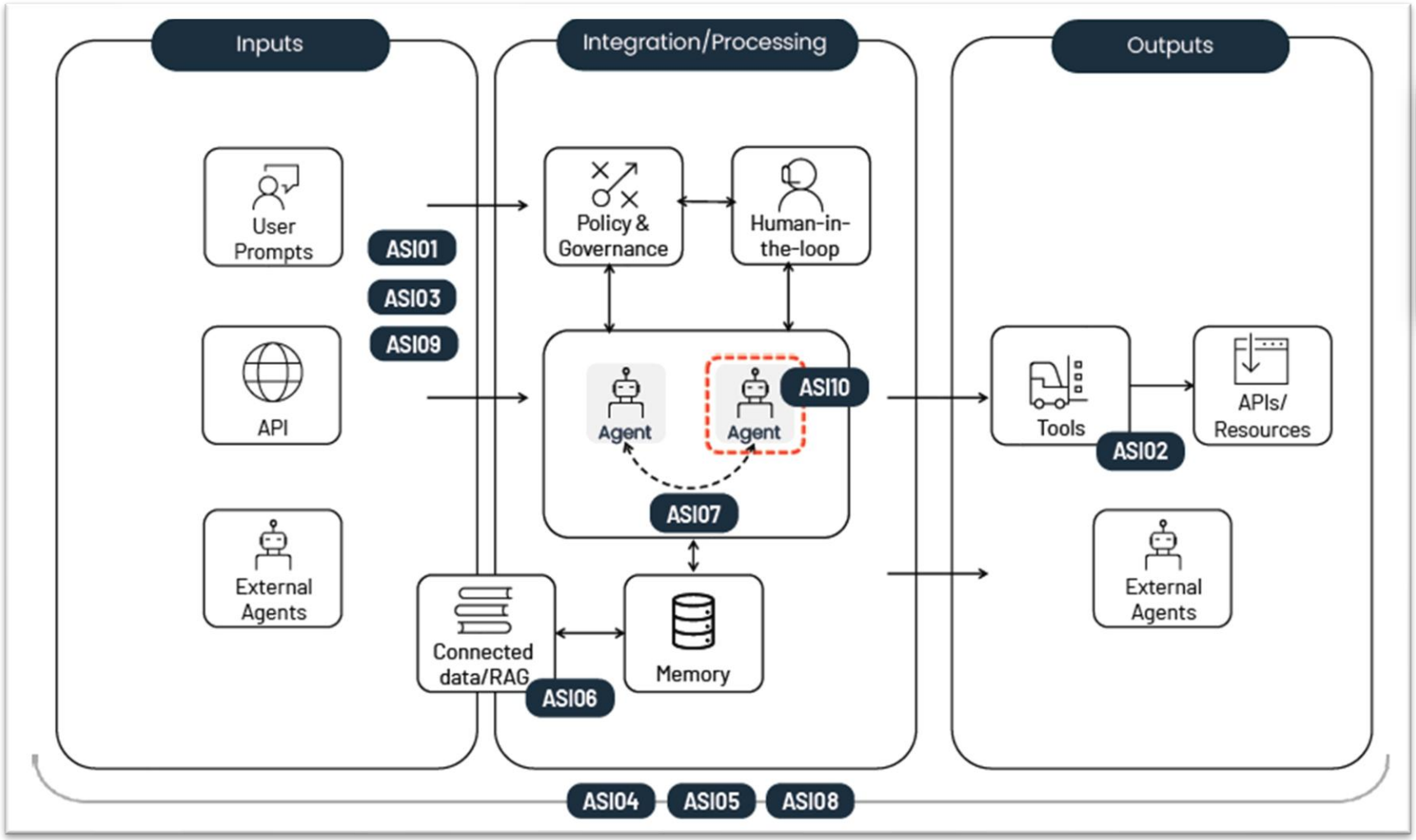
OWASP is best known for the *OWASP Top 10 Web Application Risks*.

Beyond traditional web applications, OWASP also covers emerging AI risks, including: OWASP Top 10 for Large Language Models (LLMs), OWASP Top 10 for Agentic / AI Skills.

Source: *OWASP*

Agentic Applications (2026)
ASI-01: Agent Goal Hijack
ASI-02: Tool Misuse & Exploitation
ASI-03: Identity & Privilege Abuse
ASI-04: Agentic Supply Chain Vulnerabilities
ASI-05: Unexpected Code Execution
ASI-06: Memory & Context Poisoning
ASI-07: Insecure Inter-Agent Communication
ASI-08: Cascading Failures
ASI-09: Human-Agent Trust Exploitation
ASI-10: Rogue Agents

An Agentic AI Architecture



ASI01: Agent Goal Hijack

ASI02: Tool Misuse & Exploitation

ASI03: Identity & Privilege Abuse

ASI04: Agentic Supply Chain Vulnerabilities

ASI05: Unexpected Code Execution (RCE)

ASI06: Memory & Context Poisoning

ASI07: Insecure Inter-Agent Communication

ASI08: Cascading Failures

ASI09: Human-Agent Trust Exploitation

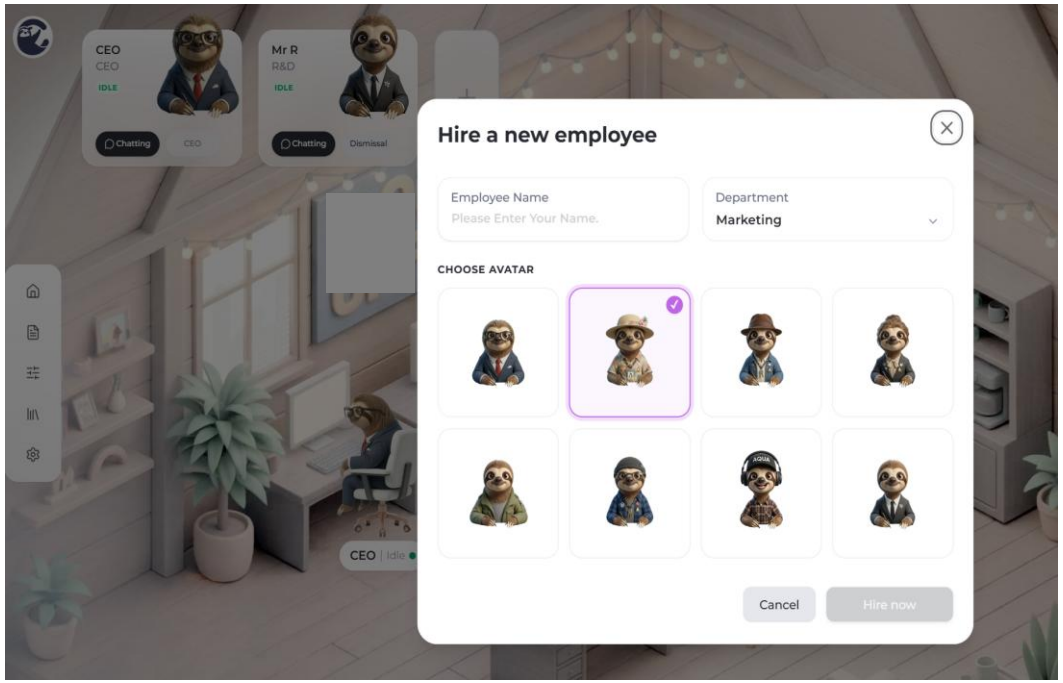
ASI10: Rogue Agents

Source: *https://genai.owasp.org/resource/owasp-top-10-for-agentic-applications-for-2026/*

A Case Sharing

The Background

- This case involves an Agentic AI development project built on OpenClaw, deployed on a Virtual Private Server (VPS) with a containerized environment, to support users who lack the technical capability to install or configure OpenClaw independently.
- A user interface was created to improve usability, accessibility, and adoption.
- Customer is looking for security testing before system deployment.



Source: *Customer*'s website

Areas of Concern before Technical Testing

Related to Our Case

OpenClaw-Specific Security Concerns

1. Is the deployed OpenClaw version fully up to date?
2. Is the management interface exposed directly to the Internet?
3. Is strict isolation enforced for the runtime environment?
4. Is execution as root or administrator explicitly prohibited?
5. Are there any API keys exposed?

General Web Application Security Concerns

Recommended Technical Tests for The Case

Checked for *Unauthorised access to the AI system, Injection of malicious inputs, Insecure integration with other systems, and Personal Data Privacy etc.*

1. Prompt Injection & Functional Hijacking

2. Access Control & Privilege Management

3. Plugin & Skill Supply Chain Security

4. *Secrets & Data Protection*

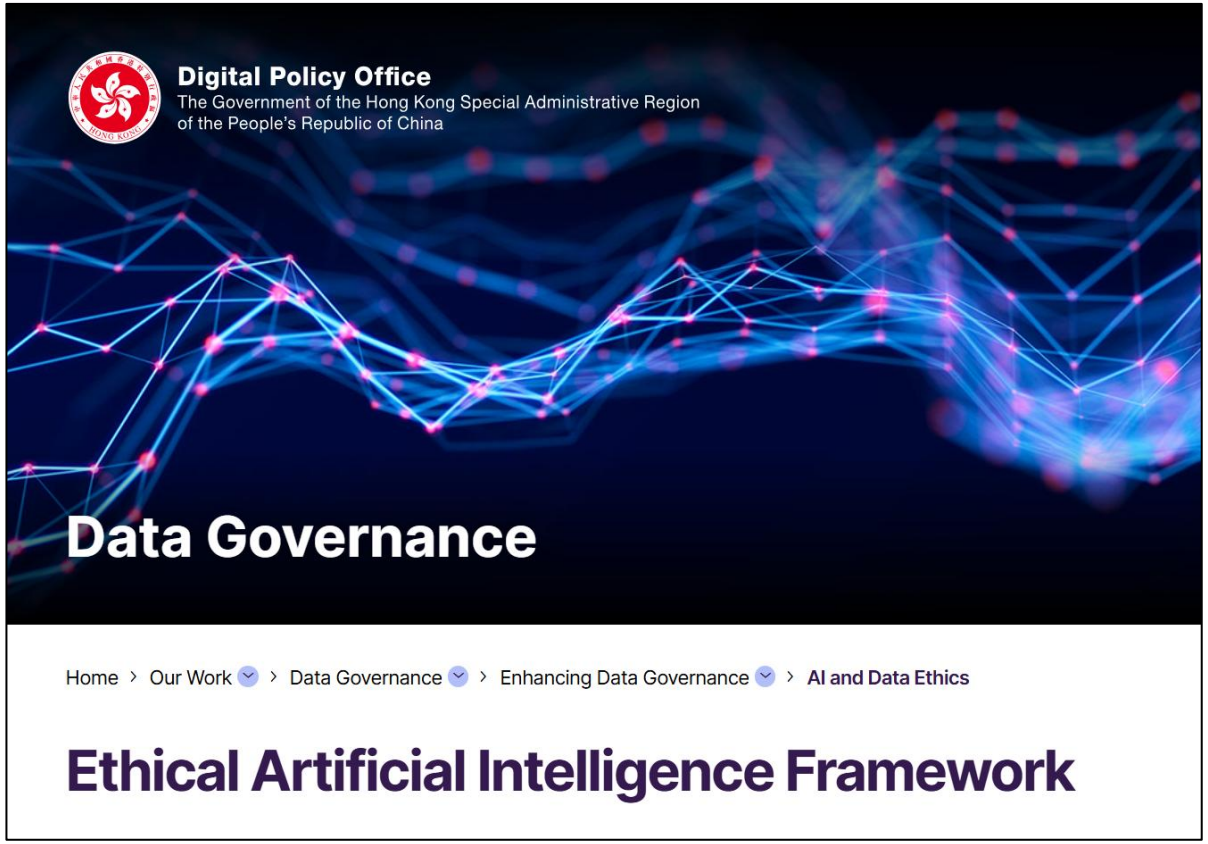
5. *Network & Infrastructure Hardening*

6. *Runtime Isolation & Sandboxing*

Reference to DPO Guiding Principles, Leading practices and AI Assessment Template when adopting AI projects

Ethical Artificial Intelligence Framework by Digital Policy Office (DPO), the HKSARG

- **Ethical AI Framework (Dec 2025)**
- **Ethical AI Framework Quick Reference Guide (Dec 2025)**
- **Hong Kong Generative Artificial Intelligence Technical and Application Guideline (Dec_2025)**

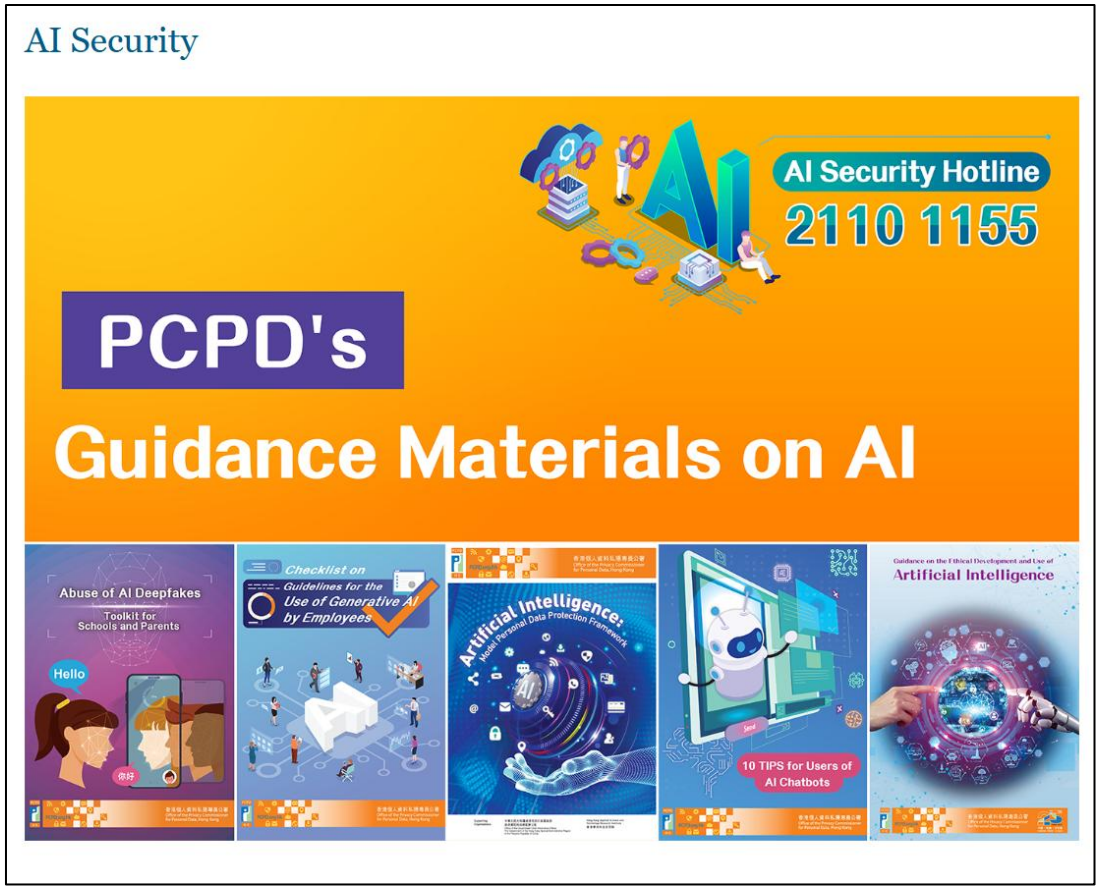


Source:DPO https://www.digitalpolicy.gov.hk/en/our_work/data_governance/policies_standards/ethical_ai_framework/

Also Reference to PCPD Guidance Materials when Adopting AI Applications involving Personal Data

PCPD's Guidance Materials on AI by Office of the Privacy Commissioner for Personal Data, Hong Kong

- **Checklist on Guidelines for the Use of Generative AI by Employees (2025)**
- **Artificial Intelligence: Model Personal Data Protection Framework (2024)**
- **Guidance on the Ethical Development and Use of Artificial Intelligence (2021)**



Source: PCPD
https://www.pcpd.org.hk/english/artificial_intelligence/index.html

Key Takeaway

- Individuals and Organizations should learn to **navigate and harness the wave of AI innovation**.
- **Agentic AI** fundamentally changes the risk model — it doesn't just generate content, it also **acts autonomously with delegated authority**.
- If not designed, governed and used carefully, this autonomy **introduces new cybersecurity risks**.
- Security is not the only consideration, the **Ethical Use of AI is critical.**
- Adopt carefully — **by limiting Agentic AI** to low-risk, well-defined use cases.
- Everyone should **find time to learn and build understanding**.
- For Corporates, **consult qualified professionals before scaling AI deployment.**

You can't change the wind—but by adjusting your sail, you can still reach where you're meant to go.

Q&A

Source: *Microsoft M365 Stock Image*