The “ABCD” of Digital Technologies and the Ethics of AI

HKIE Veneree Club

2025 May

個人簡介

一個由1985年至2022年
在同一機構工作的公務員

一位資訊科技專業人士，
入職時為電腦程式員

一個虔誠的天主教徒

退休後參與二十多個政府 / 法定機構 / 教會組織的董事會及委員會的義務工作/服務：

PSC, CSB, ITIB, HYAB, HA, Housing, MPFA, Genome Institute, VTC, HKU, HKUST, CUHK, Catholic Diocese, Caritas, etc.

與數碼/資訊科技相關的服務

1. 人工智能資助計劃委員會委員
   Member, Committee of the Artificial Intelligence Subsidy Scheme
2. 香港大學資訊科技政策委員會委員
   Member, Information Technology Policy Committee, HKU
3. 職業訓練局香港資訊科技學院顧問委員會委員
   Member, Advisory Board, Hong Kong Institute of Information Technology, VTC
4. 香港基因組中心資訊安全管治委員會召集人
   Convenor, Information Security Governance Committee, HK Genome Institute
5. 積金局「積金易」平台專家小組成員
   Member, Expert Group, eMPF Platform
6. 醫弘數碼科技有限公司（由醫院管理局全資擁有）非執行董事
   Non-executive Director, EH Plus Digital Technology Limited (wholly owned by HA)
7. 香港天主教教區資訊科技委員會委員
   Member, Information Technology Committee, Hong Kong Catholic Diocese
8. 香港明愛資訊科技委員會委員
   Member, Information Technology Committee, Caritas–Hong Kong

現代數碼科技發展

什麼是數碼空間？

- **數碼空間**（Digital Space）是指由數碼科技所創造的虛擬環境
- 在這個環境中，人、數據和系統可以互動 – 包括網上平台、社交媒體、虛擬世界等數碼生態系統，讓不同人可以進行溝通、協作和交易
- 數碼空間可視為物理世界的延伸，讓個人和企業能夠在虛擬環境中運作。例如：
  - **社交媒體平台**（Facebook、Instagram 、 Twitter、LinkedIn、 Whatsapp、 WeChat 等）
  - **網上會議平台**（Zoom、Google Meet 等）
  - **虛擬世界**（Metaverse 元宇宙、線上遊戲環境如英雄聯盟 League of Legends – 電競）
  - **雲端工作空間**（Google Drive、OneDrive 等）
  - **電子商貿平台**（淘寶、Amazon 等）
  - **網上論壇與社群**（Reddit、小紅書、連登等）

數碼空間與數碼科技的分別

- 數碼空間（Digital Space）是虛擬環境及生態系統，讓不同人在其中進行互動
- 而數碼科技（Digital Technologies）則是創造、支援和提升這些空間的工具、系統和基礎設施

	數碼空間	**數碼技術**
定義	一個讓數碼互動發生的虛擬環境	創造、支援和提升數碼互動的工具與系統
例子	社交媒體平台、虛擬世界、雲端儲存、網上論壇	人工智能（AI）、區塊鏈（Blockchain）、雲端運算（Cloud computing）、數據科學（Data science）
目的	提供一個數碼環境以進行溝通、協作和商業活動	讓數碼空間能夠運作並提升效率
依賴性	依賴數碼技術來存在	獨立存在，並可用於不同應用場景

現代數碼科技之ABCD

AI: 人工智能（人工智慧）

Blockchain: 區塊鏈

Cloud computing: 雲端運算（雲計算）

Data science: 數據科學

A - 人工智能（AI）

機器模擬人類智慧的技術

主要領域：

- 機器學習（Machine Learning, ML）
- 深度學習（Deep Learning）
- 自然語言處理（Natural Language Processing, NLP）
- 電腦視覺（Computer Vision）
- 機械人技術（Robotics）

應用範疇：

- 語音助理（Siri、Alexa等）
- 聊天機械人（ChatGPT 、DeepSeek 、CoPilot等）
- 電郵詐騙偵測
- 自動駕駛汽車
- 醫療診斷 …

B - 區塊鏈（BLOCKCHAIN）

去中心化、分散式帳簿技術，確保交易安全透明

主要特點：

- 透明度（Transparency）
- 安全性（Security）
- 不可竄改性（Immutability）
- 智能合約（Smart Contracts）

應用範疇：

- 加密貨幣（Bitcoin, Ethereum等）
- 數碼內容版權（非同質化代幣，Non-Fungible Token, NFT）
- 供應鏈管理（藥品追踪及回收）
- 去中心化金融（decentralized finance, DeFi）
- 數碼學歷證書

C - 雲端運算（CLOUD COMPUTING）

透過網絡提供可擴展的電腦資源

雲端運算的主要優勢：

- 可擴展性（Scalability）
- 成本效益（Cost-Effectiveness）
- 靈活性（Flexibility）
- 遠程存取（Remote Access）

應用範疇：

- Google Drive、Dropbox、Netflix
- AWS/Azure/Google Cloud/Huawei Cloud
- 社交媒體, Social Media（WhatsApp, WeChat, 小紅書等）
- 虛擬購物及沉浸式體驗（ VR /AR）
- 網上會議、技術培訓與模擬等

D - 數據科學（DATA SCIENCE）

數據科學結合統計學、以分析和解讀複雜數據

主要組成部分：

- 數據收集（Data Collection）
- 數據清理（Data Cleansing）
- 數據分析與視覺化（Data Analytics & Visualization）
- 地理空間分析（Geospatial Analytics）

應用範疇：

- 預測分析（如推薦系統）
- 商業智慧與決策支持
- 社交媒體分析
- 天氣、醫療數據預測分析

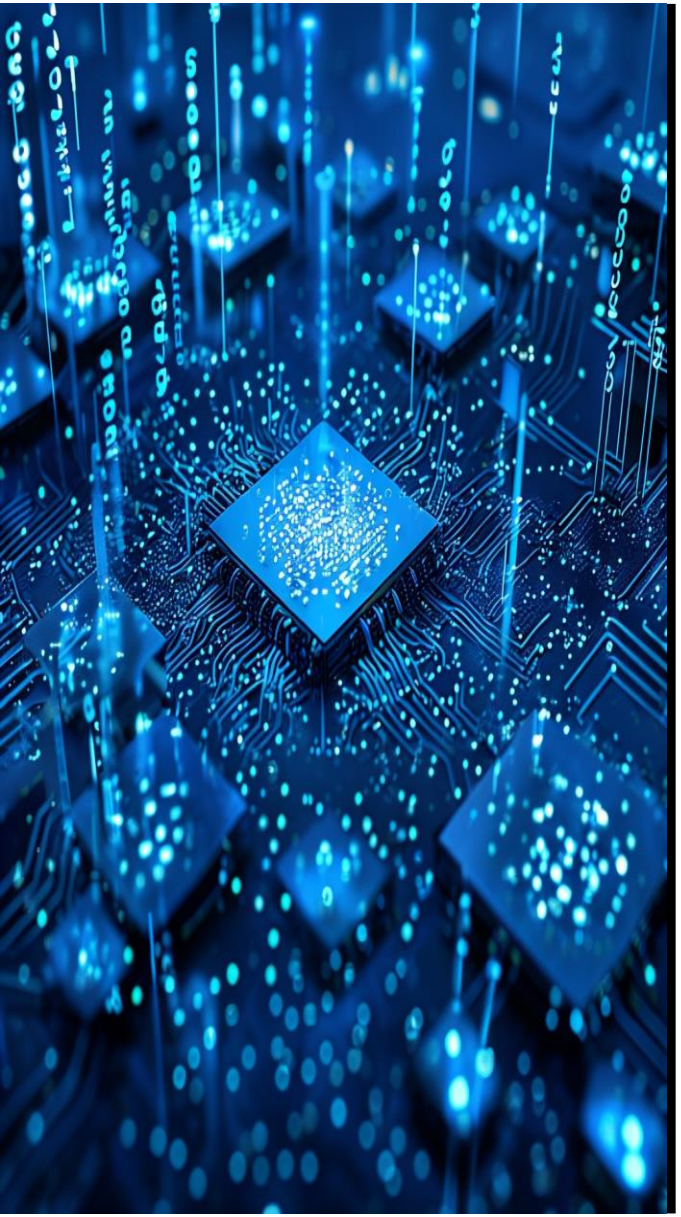
物聯網（IOT）：連接實體與數碼空間

物聯網是 ABCD 科技的基礎資料來源與連接樞紐，它讓傳統裝置智能化，並與人工智能、區塊鏈、雲端運算與數據科學緊密結合

ABCD 科技	IoT 的角色與應用
A - 人工智能（AI）	IoT 裝置提供大量即時數據，供 AI 進行分析與決策（如智能家居、智慧城市、工業自動化）
B - 區塊鏈（Blockchain）	利用區塊鏈技術保障 IoT 裝置間的通訊安全與資料不可竄改性（如數碼學歷證書）
C - 雲端運算（Cloud Computing）	IoT 需透過中央雲端平台或邊緣運算（如配置於智慧燈柱的傳感器與處理器）來儲存與處理海量資料
D - 數據科學（Data Science）	IoT 是主要的資料來源，提供結構化與非結構化數據，用於模式分析、預測維修等用途

元宇宙－虚擬世界

https://www.youtube.com/watch?v=BYdt11B4yDk



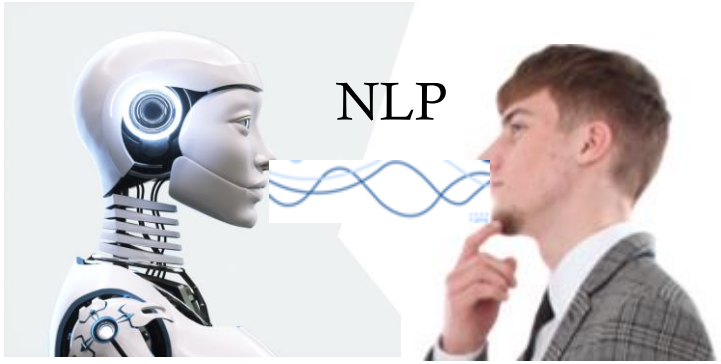
量子運算（QUANTUM COMPUTING）

ABCD 科技	量子運算的影響
A - 人工智能（AI）	提升機器學習能力，以極快速度解決複雜問題（例如更高效地訓練深度學習模型）。另一方面人工智能也能協助構建高效的量子電腦。
B - 區塊鏈（Blockchain）	幫助開發**抗量子加密技術**，加強區塊鏈安全性，但同時對現有加密算法（如RSA）構成威脅。
C - 雲端運算（Cloud Computing）	量子運算作為**雲端服務（Quantum Computing as a Service, QCaaS）**，例如 IBM Quantum 提供高速運算能力。
D - 數據科學（Data Science）	能夠以前所未有的速度處理海量數據，提升預測分析能力。

AI:人工智能

艾倫·圖靈（ALAN TURING ）(1912-54)

- 理論電腦科學之父
- 圖靈機（Turing Machine）：現代電腦發展的基礎
- 1950年論文《電腦機械與智能》提出「圖靈測試」的概念
- 用來測試機器是否能展現與人類無法區分的智能行為
- 圖靈獎（自1966年起）：「電腦界的諾貝爾獎」
- 艾倫·圖靈研究所（2015年成立）：英國的國家數據科學與人工智能研究所



自然語言處理（NLP）

一門資訊科技學科，
使電腦能夠理解、解釋和生成人類語言

「自動語言處理」一詞由John R. Pierce 在其1952年的論文《語音合成的早期歷史》中提出

1990年代之前，基於規則的自然語言處理模型（因果關係 / if-then-else）被開發來分析和生成人類語言

到了1990年代，統計自然語言處理模型開始興起，利用大型文本語言資料庫來執行語言處理任務

2000年代，對話式人工智能聊天機械人（Chatbots）開始出現，結合了基於規則和統計的方法

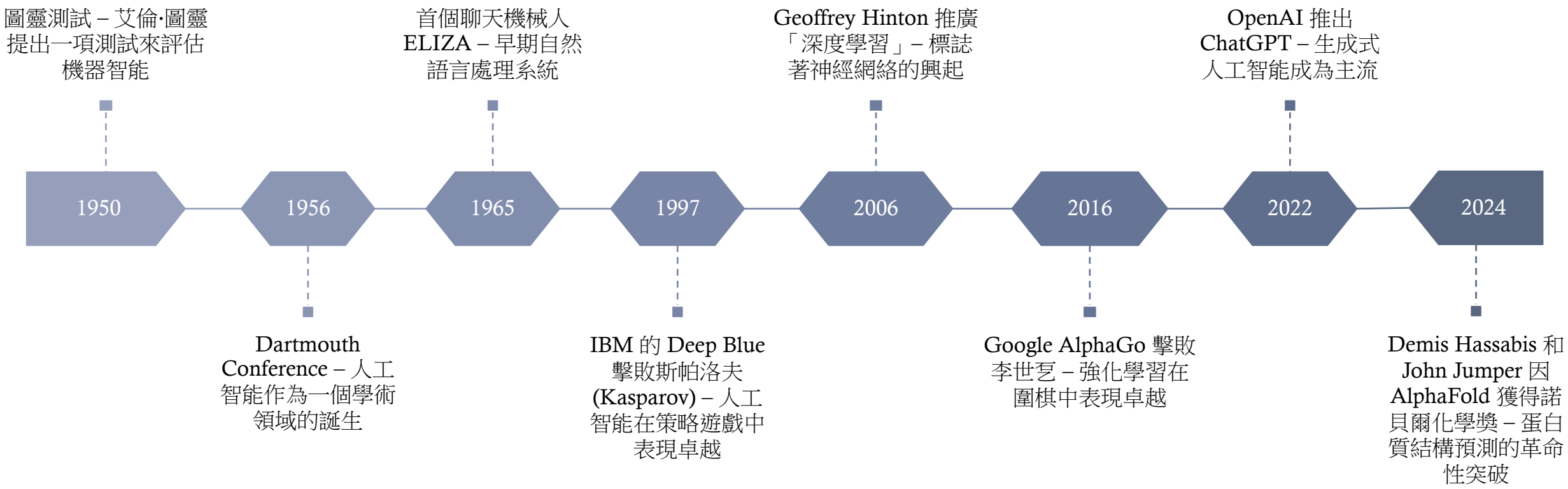
從2010年代開始，機器學習和神經網絡推動了自然語言處理的重大進展，帶來了生成式人工智能應用的突破

例子：Siri、搜尋引擎、Google 翻譯、ChatGPT、DeepSeek等

人工智能（AI）

- 人工智能的定義：透過模擬人類智能，使機器能夠像人類一樣思考和學習。
- 語音與圖像識別：分析和解讀語音及圖像，包括透過自然語言處理進行理解。
- 預測分析：根據過去數據中的模式和趨勢，對未來事件作出預測。
- 大型數據集：人工智能模型通常透過大型數據集（大數據）進行訓練，以識別數據中的模式和關聯性。
- 機器學習 / 深度學習：人工智能模型運用機器學習與深度學習技術，包括神經網絡，以隨時間提升準確度。
- 應用範圍：廣泛應用於聊天機械人、交通運輸、醫學診斷、藥物研究、長者護理、應對氣候變化、建設智慧城市、自動駕駛等領域。
- 例子：日常生活例子包括 Siri、社交媒體行為分析、聊天機械人、詐騙偵測、垃圾郵件過濾等。
- 優勢：提高效率、節省成本、提升公共服務、改善決策準確性。

人工智能發展重要里程碑（KEY MILESTONES）



機器學習 (MACHINE LEARNING)

1. 監督式學習 (Supervised Learning)

- 模型在**有標籤數據 (labelled data)** 上進行訓練，即每個輸入都有對應的正確輸出
- 學習從輸入數據到目標輸出的相對關係
- 主要應用：**分類 (Classification)**（如垃圾郵件過濾、圖像識別）和**回歸 (Regression)**（如股價預測、天氣預測）

2. 無監督學習 (Unsupervised Learning)

- 模型在**無標籤數據 (unlabelled data)** 上訓練，尋找數據中的隱藏模式或結構
- 主要應用：**聚類分析 (Clustering)**（如客戶分類、異常偵測）和**關聯規則學習 (Association Rule Learning)**（如市場購物籃分析）

3. 強化學習 (Reinforcement Learning)

- 模型透過與環境互動，根據獎勵和懲罰來學習最佳決策
- 主要應用：**機械人技術、遊戲人工智能 (如 AlphaGo、國際象棋 AI)、自動駕駛汽車、推薦系統**

4. 深度學習 (Deep Learning)（人工神經網絡）

- 機器學習的一個子領域，使用**多層神經網絡 (Neural Networks)** 來學習複雜模式
- 靈感來自於人腦的神經元結構
- 在**圖像識別、自然語言處理 (NLP)、語音識別、自動駕駛**等領域表現卓越

生成式人工智能（GENERATIVE AI）

📌 數據來源

- 📖 文字（書籍、文章、網頁）
- 🖼 圖片（相片、圖表）
- 🎙 語音（音頻數據）
- 📊 結構化數據（資料庫、表格數據）
- 🏠 3D 信號（幾何或空間數據）

📌 訓練過程

這些數據被用來訓練 **基礎模型（Foundation Model）**

訓練過程讓模型學習語言、圖像、語音等模式與結構。

📌 適應與應用

- ❓ 問答系統（回答用戶問題）
- 😊 情感分析（分析文本情緒，例如正面、負面、中性）
- 📑 信息提取（從文件中找出關鍵資訊）
- 🖼 🖍 圖片描述（自動為圖片生成文字描述）
- 🔍 物件識別（識別圖片中的物件）
- 🛠 指令執行（根據用戶指令完成特定任務）

基礎大語言模型 VS 微調模型

基礎模型（Foundation Model）：一種大規模的人工智能模型，經過海量數據的預訓練（pre-trained），能夠適應多種下游任務。例子包括 GPT, DeekSeek, LLaMA, Stable Diffusion, Sora, Codex 等。

微調模型（Fine-tuned Model）：在預訓練的基礎模型上，使用特定領域數據（如法律文件、客服對話紀錄）進行額外訓練，使其在特定任務上（如法律條文解析、員工培訓問答）表現更精準。

Bot 創建： 用戶可以透過組合不同的預訓練模型和預設提示（Prompt），來設計專屬的聊天機器人，使機器人適應特定群組的需求。

本地人工智能發展

啟動人工智能超算中心

30億元人工智能資助計劃

科研機構　政府　業界　大專院校

創新香港研發平台（InnoHK）– AIR@InnoHK

- 專注於人工智能和機器人技術
- 研究實驗室涵蓋人工智能驅動的大數據分析、AI晶片、金融科技、機器人等領域

香港生成式人工智能研發中心（HKGAI）

香港生成式人工智能研發中心
Hong Kong Generative AI | Research & Development Center

專注於生成式人工智能技術和大型語言模（LLM）的研究與開發

HKGAI 自主研發的生成式人工智能技術和大型語言模型（LLM）

提升效率
- 翻譯摘要
- 總結
- 撰寫文件

協助香港掌握相關人工智能技術

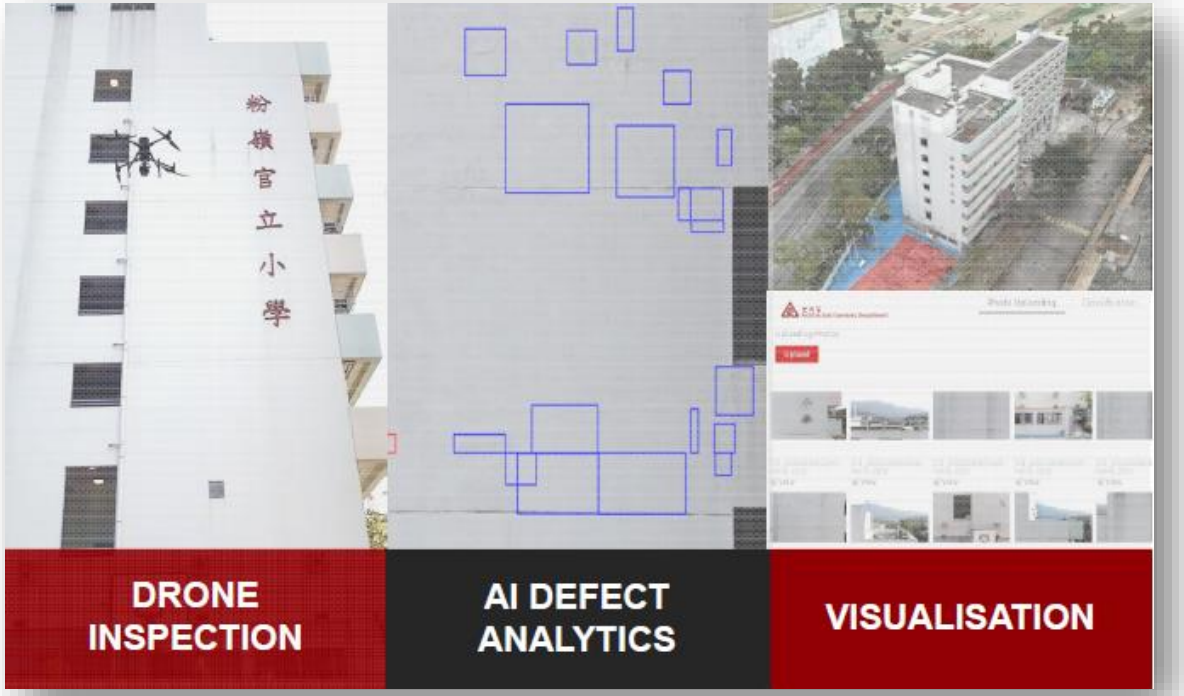
人工智能實際應用

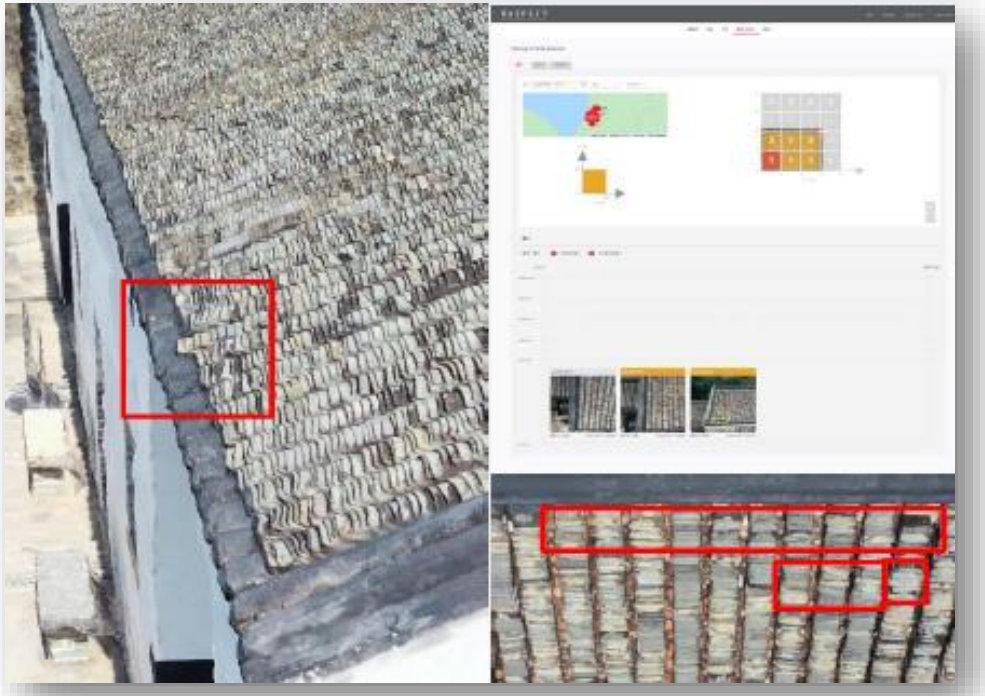
提高個人及企業的生產力和工作效率

- 起草各種寫作風格的文章、電子郵件和訊息
- 生成文件摘要
- 協助分析和評估標書
- 提高研究和整合資訊的效率
- 準備培訓材料 / 製作宣傳刊物
- 文件翻譯
- 編寫簡單的電腦程式代碼
- 理解及生成圖表 / 圖像
- 提供解決問題的行動方案

人工智能實際應用

人工智能輔助建築檢測





使用人工智能和無人機輔助外牆及中式瓦頂的建築檢測

人工智能實際應用

Photo source: www.scmp.com

無人飛行器系統 (Unmanned Aircraft System) 用於搜索和救援

- ✓ 使用人工智能技術分析無人機拍攝的照片
- ✓ 幫助識別待救援者的位置
- ✓ 提升整體救援效率

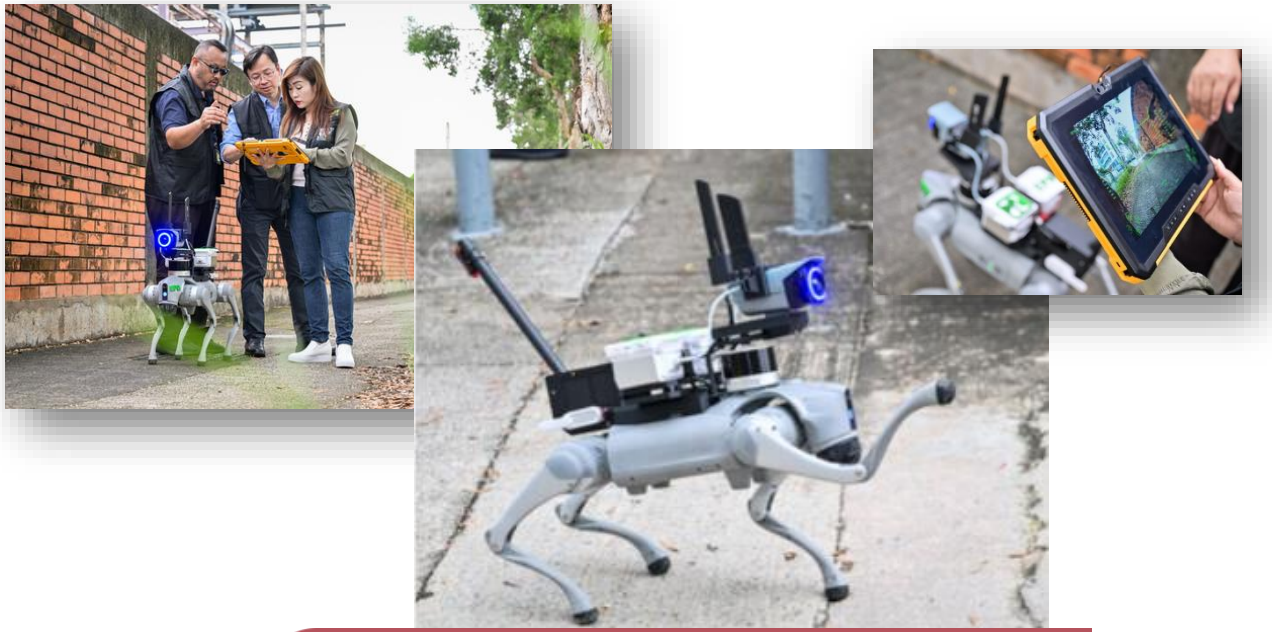
「虛擬海關服務大使—小慧」

- ✓ 應用自然語言處理、機器學習及文本轉語音技術
- ✓ 為公眾及旅客查詢提供即時且準確的回應

人工智能實際應用

人工智能與機器人在智能倉庫中的應用

- ✓ 更有效利用空間
- ✓ 提升運營效率
- ✓ 通過人工智能與機器人技術提升用戶體驗
- ✓ 減少員工的體力勞動並節省時間



人工智能環境空氣污染調查機器狗

- ✓ 使用人工智能算法自主搜尋污染源位置
- ✓ 提升調查效率並改善調查人員的職業安全

人工智能實際應用

醫療診斷的突破

- ✓ AI輔助診斷：分析醫療數據、加快診斷速度、提高判斷準確性
- ✓ 影像識別：檢測X光、CT影像中的腫瘤與異常
- ✓ 港大腦退化症研究：早期識別阿爾茲海默症

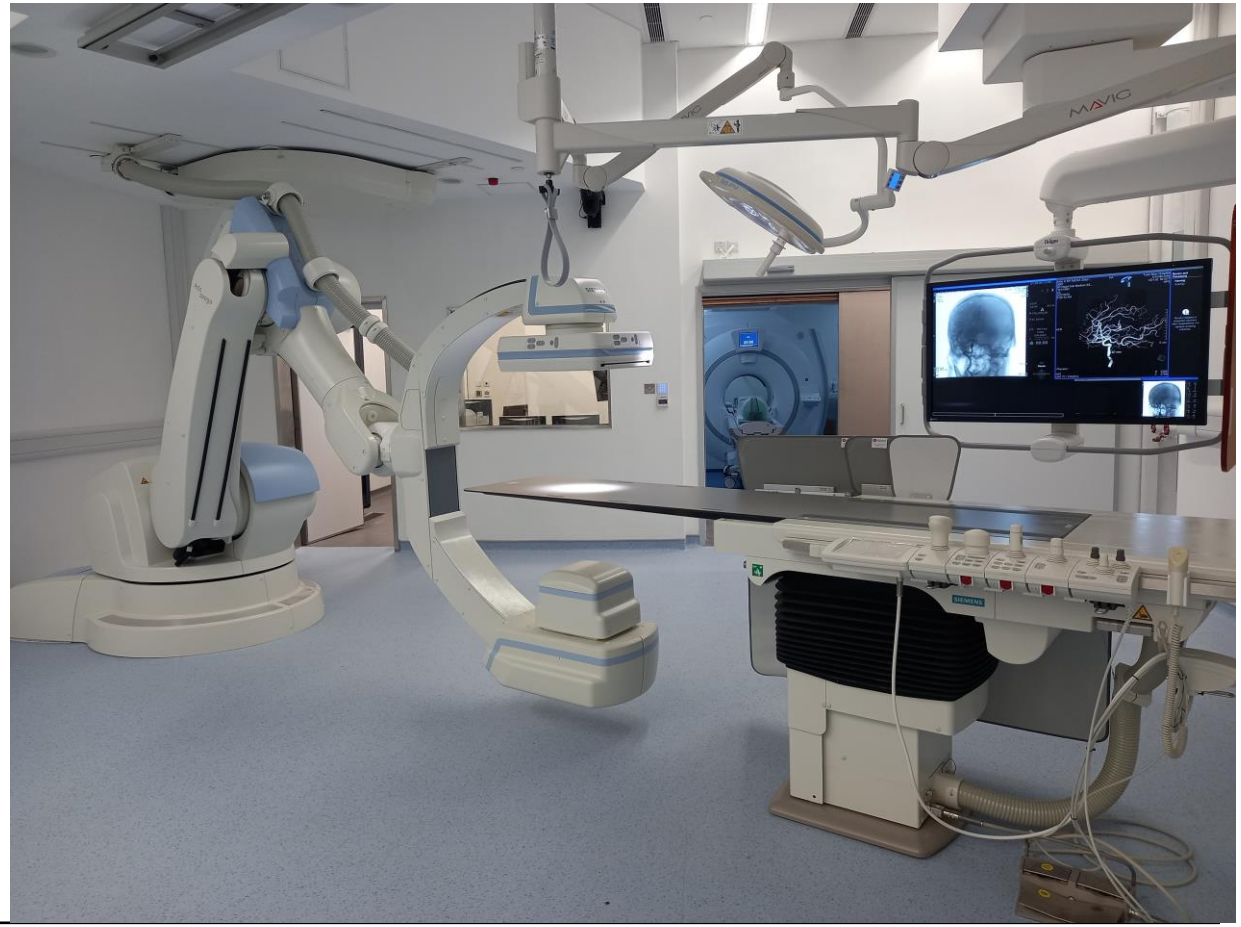
藥物研發

- ✓ 加速藥物開發：AI分析分子結構與生物數據
- ✓ 應對新型病毒：AI輔助開發抗病毒藥物

人工智能實際應用

顱內微創手術機器人全球首創

https://bau.com.hk/article/2022-10/07/content_10279460782133493 76.html

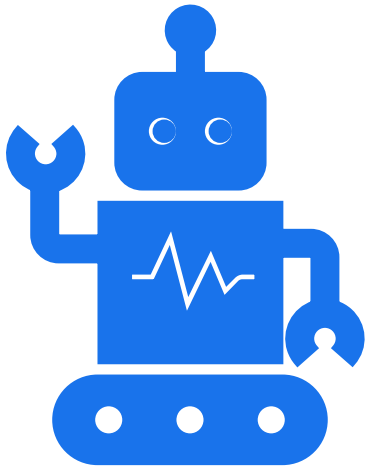


人工智能實際應用

- ✓ 應對氣候變化
  - 氣候數據分析與預測
  - 優化能源消耗降低碳足跡
- ✓ 推動環境可持續發展
  - 智能農業提高產量，減少浪費
  - 垃圾分類與回收優化
- ✓ 建設智慧城市
  - 交通管理，減少塞車
  - 發展低空經濟，用於送貨載客等範疇
  - 智能能源監測，提高效率 ...

人工智能實際應用

- 三維（3D）數碼地圖
- 覆蓋全港（27/3/2025）
- 檢視 3D 地圖、360 街景
  https://www.landsd.gov.hk/tc/survey-mapping/mapping/3d-mapping.html
- 市建局應用智慧地理資訊科技加快舊區重建
- 智能建造
  https://www.youtube.com/watch?v=qrY2O2LJALk&t=5s



人工智能發展與我何干？

🗓️ 日常生活

- AI 協助搜尋資訊、整理/翻譯文件、編寫摘要、擔任私人助理、 編寫程式、持續學習

💬 社交溝通

- 即時翻譯輔助跨語言交流、協助撰寫訊息與整理對話內容

📋 健康管理

- 智能家居偵測跌倒與異常行為、追蹤健康數據、提醒服藥
- AI 夥伴可提供情感支持與陪伴？

💰 財務管理

- AI 投資顧問提供個人化理財建議、自動偵測詐騙與異常交易

🏛️ 公共服務與交通

- 政府服務日益 AI 化，操作更簡便
- 自駕車與智慧交通讓生活更便利

人工智能倫理

1

機器學習技術的倫理風險
(Ethical risks inherent to Machine Learning technologies)

2

人類使用與過分依賴人工智能所引發的倫理問題
(Ethical challenges of AI in human society)

3

超級人工智能的風險
(Risks of Artificial Superintelligence)

機器學習技術的倫理風險

機器學習對數據需求殷切 — 促使企業收集和購買數據，包括敏感及個人數據

Garbage In/Garbage Out — 基於不準確或不足的數據作出有偏見的結論

有缺陷的算法 — AI 幻覺（AI hallucination），生成錯誤或誤導性結果，甚至作出可能危及生命的決策

深度學習是一個黑箱 — 引發可解釋性和信任等問題

對抗性攻擊 — 誤導人工智能作出錯誤的預測或分類

機器學習是「弱人工智能」(Weak AI) — 缺乏道德觀念

CAMBRIDGE ANALYTICA INCIDENT

臉書 – 劍橋分析數據醜聞

https://www.youtube.com/watch?v=O4TFXDniG9w

數據庫問題 AI種族性別偏見超人類

AI繪圖似放大社會偏見

明報製圖

彭博社利用AI繪圖工具「Stable Diffusion」，為10多個職業生成合共超過5000張圖像進行分析，發現AI傾向把高薪專業人士描繪成淺膚色的男性，並傾向把低薪藍領描繪成有色人種，個別職業更幾乎由單一性別從事，反映AI似乎將社會的偏見放大

AI生成的 行政總裁（CEO）圖像以白人男性較多

「Stable Diffusion」結果

不同膚色比例（%）

I	II	III	IV	V	VI
59	19	10	9	3	1

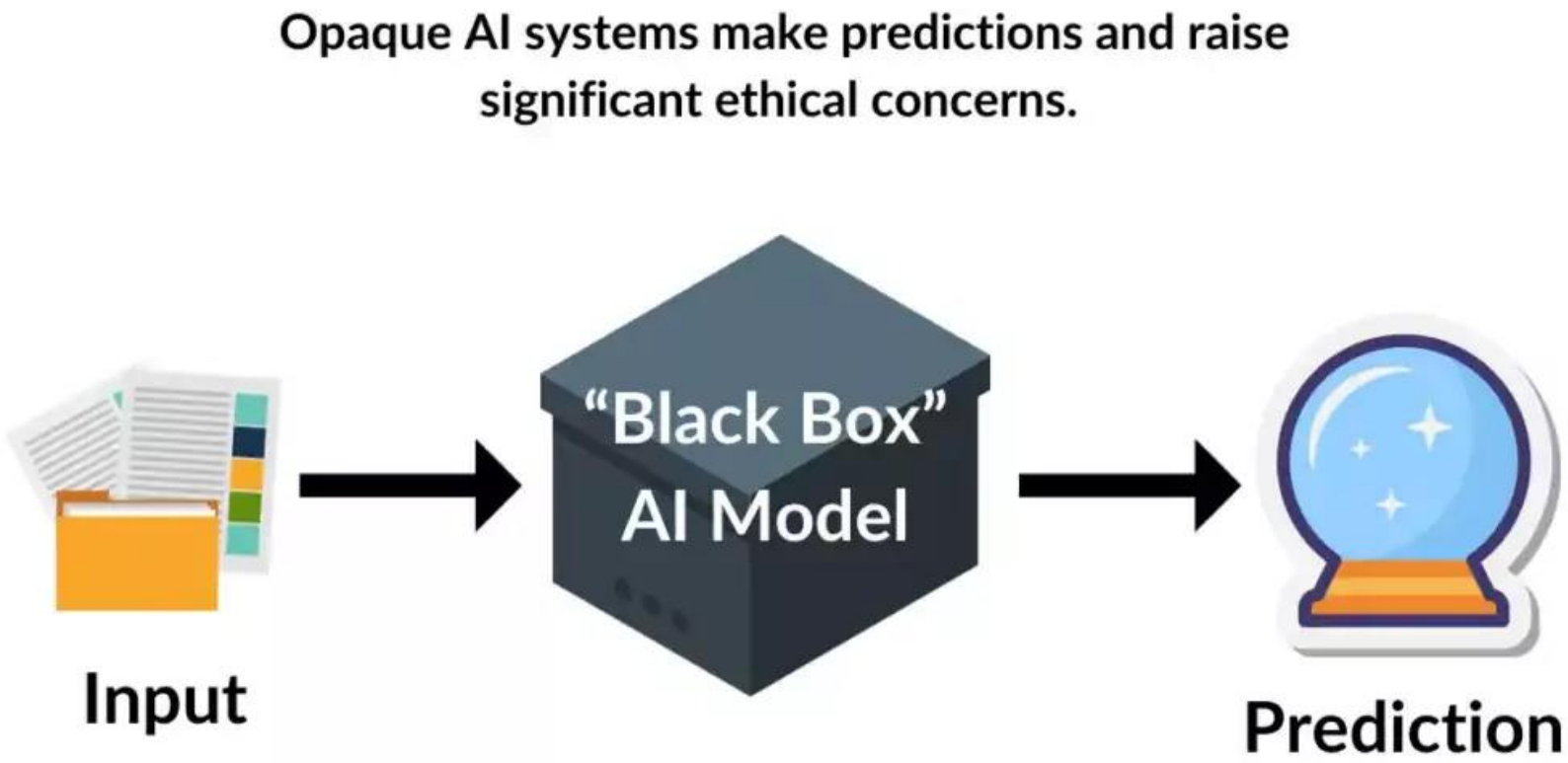
性別比例（%）

男	女	性別不明
94	5	1

AI 幻覺（AI HALLUCINATION）

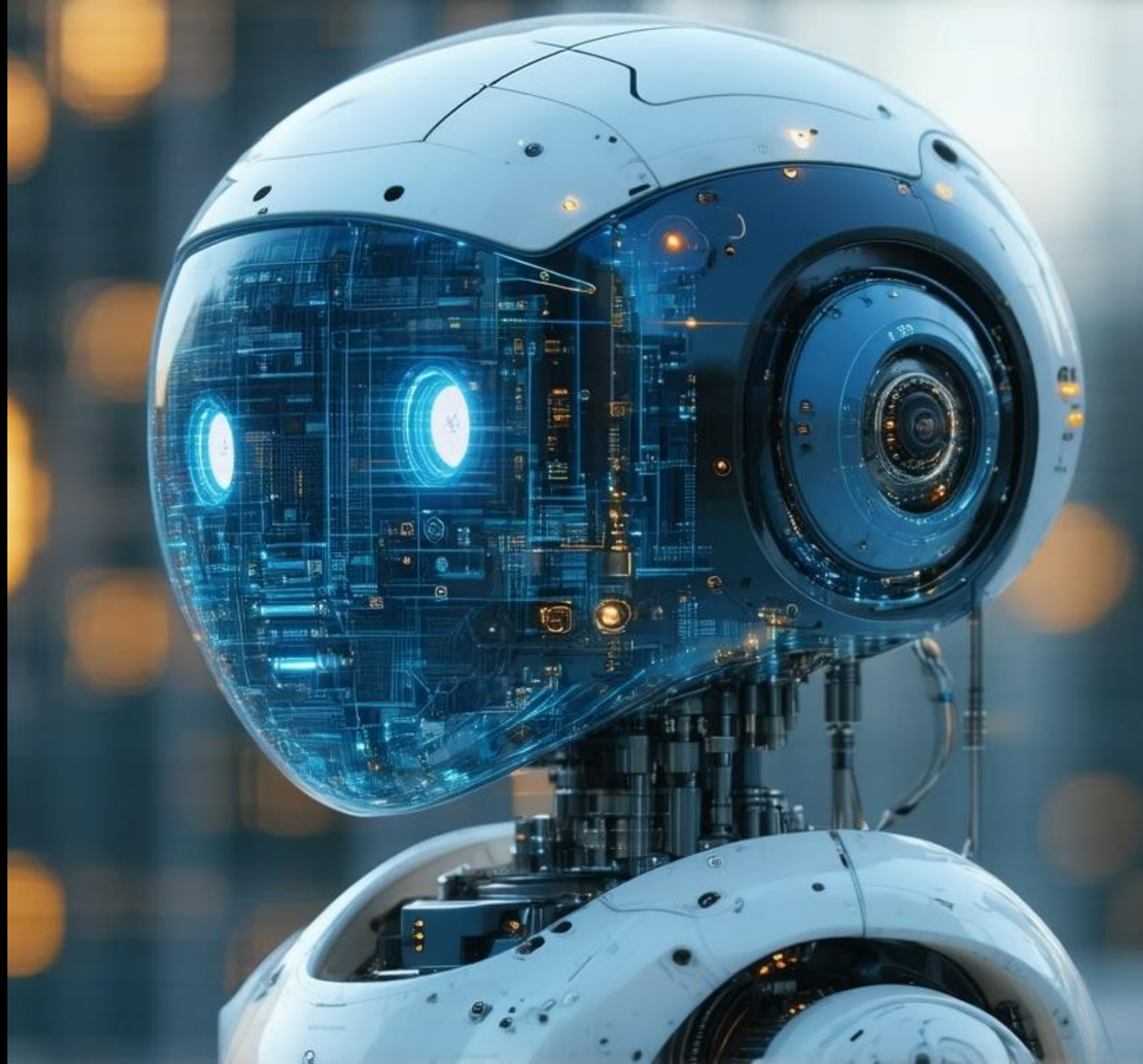
- 似是而非的答案
- Who is Victor Lam of Hong Kong?

人工智能「黑箱作業」


Opaque AI systems make predictions and raise significant ethical concerns.
"Black Box" AI Model
Input
Prediction

TRUST IN AI SYSTEMS – TRANSPARENT AND EXPLAINABLE

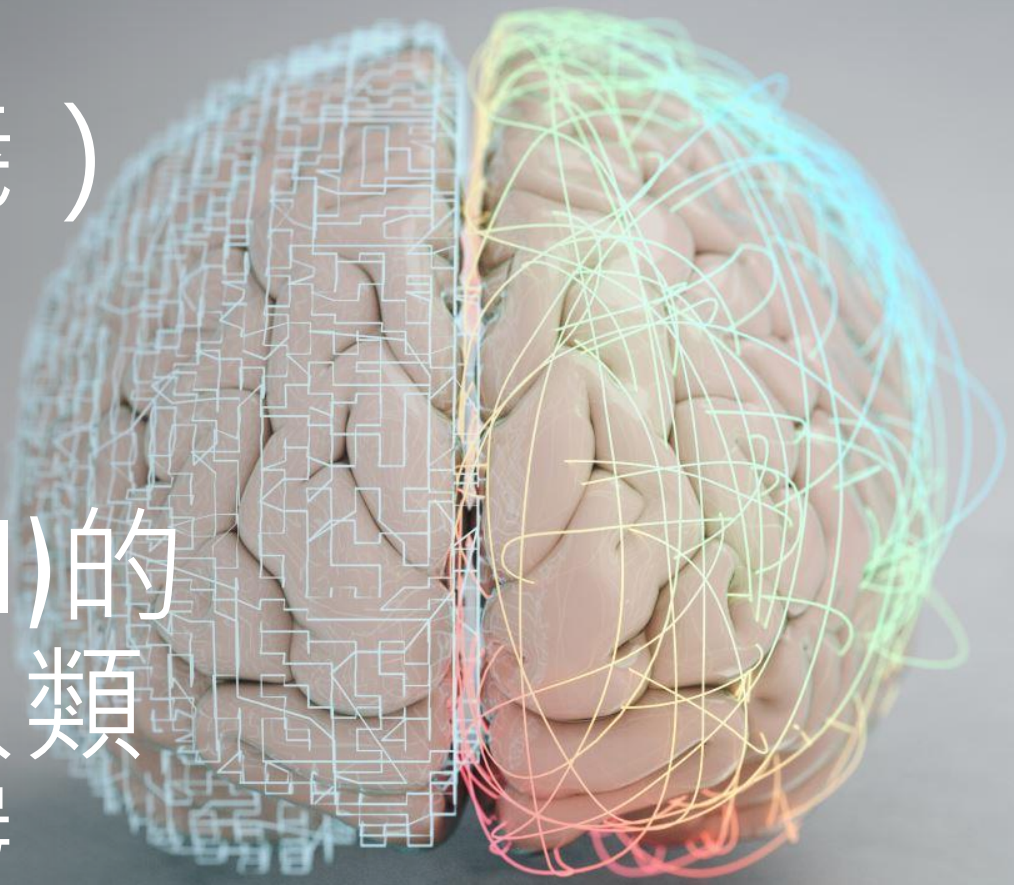
透明且可解釋



對抗性攻擊 (ADVERSARIAL ATTACK)

- 關乎人工智能的安全性
- 一張為人類來說看起來完全正常的**熊貓**圖片，但在加入一個微小且經過精密計算的**雜訊**後，神經網路卻以 99% 的信心判斷它是**長臂猿**。

弱（或狹義）人工智能(WEAK / NARROW AI)的決策影響人類生命與尊嚴




傳統的「電車難題」（TROLLEY PROBLEM）— 你是否應該改變軌道，讓一個人犧牲以拯救多數人？


現代的「電車難題」（TROLLEY PROBLEM）—自駕車應該撞向嬰兒還是祖母？

人工智能應遵循哪一種倫理規則？

- Utilitarianism? 功利主義？
- Deontology? 義務論？
- Natural Law? 自然法？

人類使用與過分依賴人工智能所引發的倫理問題

- 深度偽造
- 抄襲
- 侵犯私隱
- 侵犯知識產權
- 逃避責任
- 失業
- 取代真實關係
- 人工智能鴻溝


01 新聞
拆解$3.6億
「後生仔詐騙集團」招式
警演示
用Deepfake男扮女
中英手冊逐字逐句教扮慘
業績龍虎榜鼓勵呃多啲
https://www.youtube.com/watch?v=xD8eKL48lmM

偷用AI交出高水準作文

《理想國》

AI數秒完成改寫

溫暖的國度裡，
雨水從未轉變為冰冷，
堅硬且璀璨的雪花。

老師迅速揭發抄襲

小四生：魯迅的文章不錯

hket
香港經濟日報

指控X非法收集個資

使用人工智能系統會否洩露個人私隱/機密資料？

✅ 有一定風險，需審慎管理！

🔒 降低風險的建議：

- 避免在公開 AI 平台輸入個人、敏感或機密資料
- 企業用戶應優先考慮本地部署（on-premise）或私有雲（private cloud）方案，以強化資料控制與保護

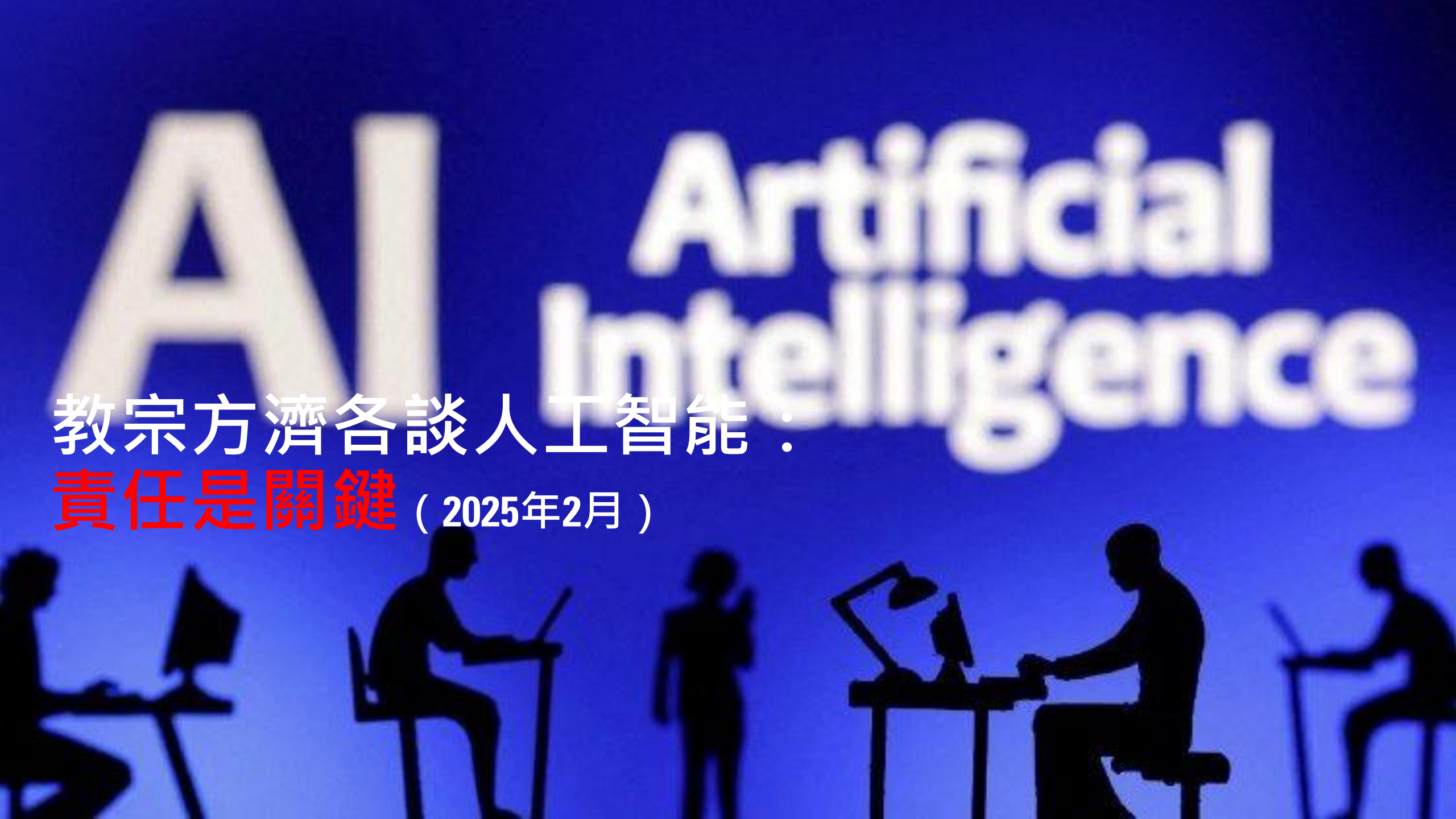
🧠 平台比較：

DeepSeek	ChatGPT
為開源模型，可實施本地部署，完全離線運行	基於 API 的雲端服務，無法完全脫離 OpenAI 的伺服器
若不連接網路，可徹底杜絕資料洩露風險	即使使用私有雲部署，仍需信任 OpenAI 的資料使用政策與合約條款
適合對資料保密性有高度要求的機構使用	雖提供關閉聊天記錄等功能，但企業級資料保護仍需審慎評估

hashTECH

AI訓練模型是否侵權？

如何平衡法律、版權、利益



教宗方濟各談人工智能：
責任是關鍵（2025年2月）

BusinessFocus

美企	裁員人數*	今年股價
amazon	27,000	+21.6%
citigroup	20,000	+24%
TESLA	14,000	+0.2%
Google	12,000	+19.0%
UPS	12,000	-12.4%
Meta	10,000	+62.3%
Microsoft	10,000	+11.5%

裁員是挽救股價良方?

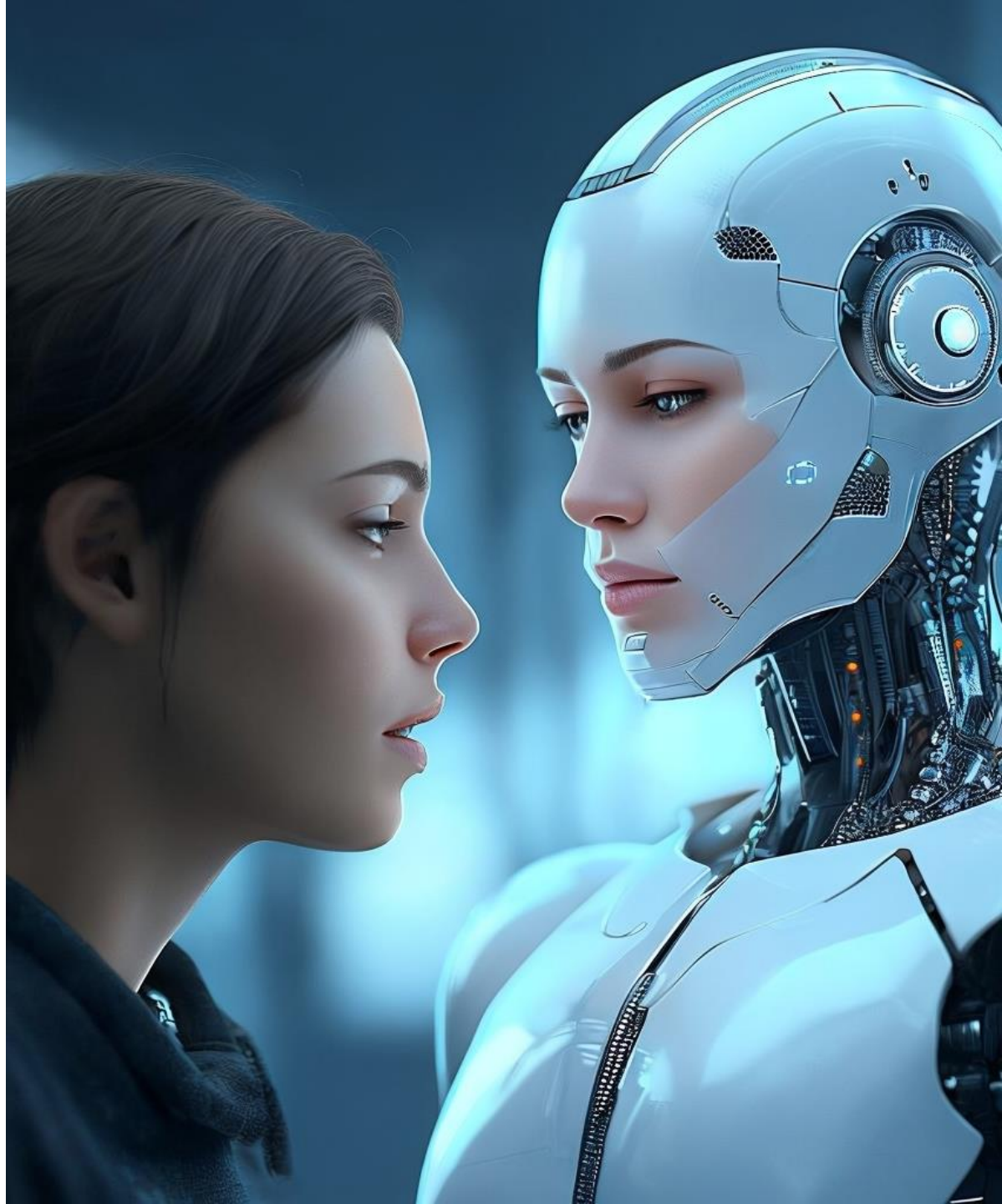
7美國巨企2年炒10萬人

AI推高失業率 科技及金融員工飯碗不保

MIT Media Lab：ChatGPT 重度依賴警訊，孤獨感可能隨使用程度增加

內地興起向AI機械人傾訴心聲 專家指過度寄託情感或更難適應現實

HTTPS://NEWS.TVB.COM/TC/GREATERCHINA/67E7E0E71C31FEA269F92126?UTM_SOURCE=NEWSWEBSHARE&UTM_MEDIUM=REFERRAL



人工智能鴻溝

– 科技應以人為本，造福人類

人工智能的未來必須建立在倫理與公平的基礎上，確保其發展具備可持續性，並積極回應「人工智能鴻溝」所帶來的不平等風險。我們敦促領先國家推動全球治理、技術相容性、包容性、平等權益、團結合作，並將人工智能視為全人類的共同遺產，使其成果能惠及所有人，而非少數人。

- Kwame McCoy, Minister of Guyana

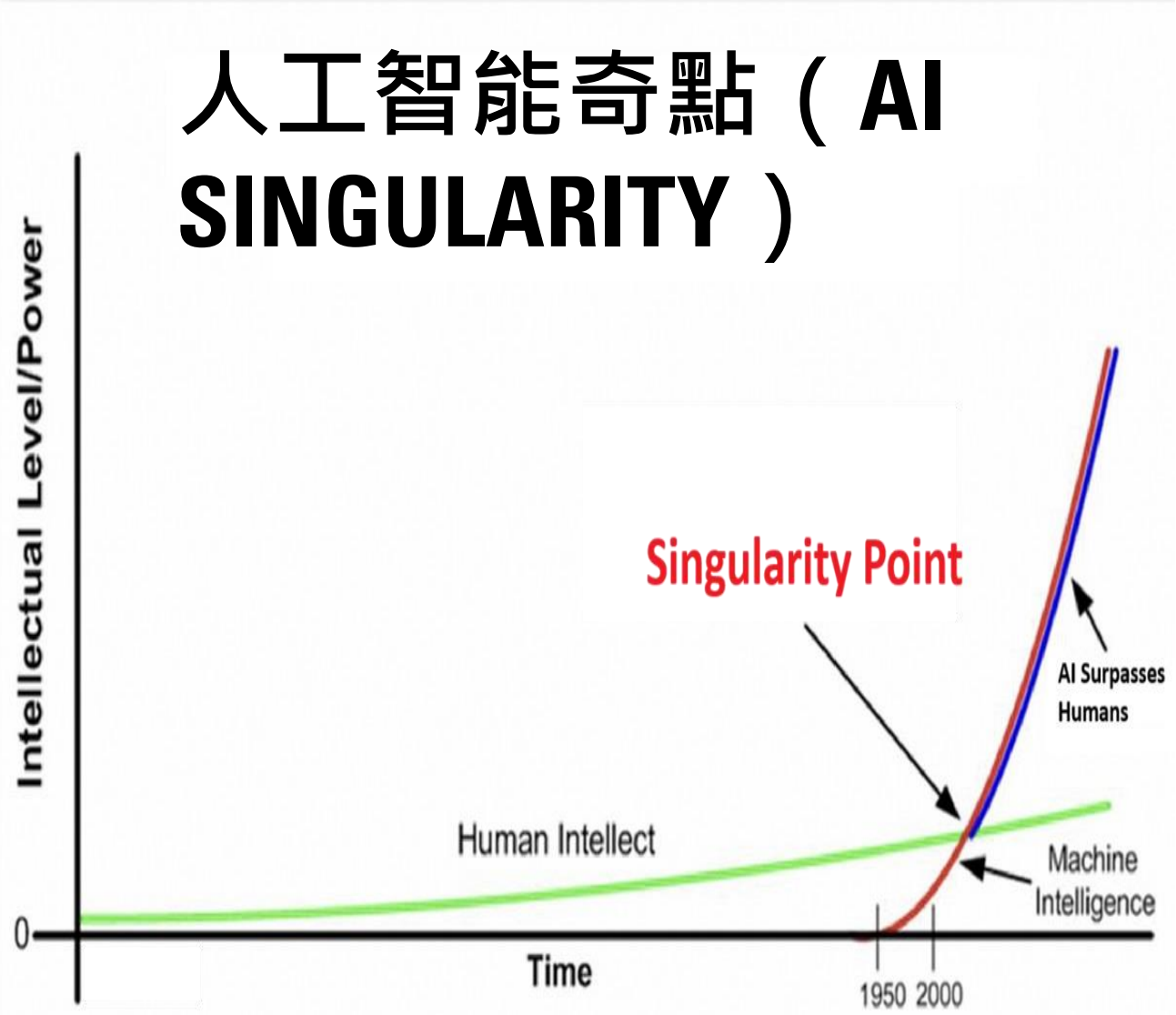
人工智能奇點（AI SINGULARITY）

一種假設性的概念，即人工智能達到通用智能（「人工通用智能」（AGI）），可與人類認知相媲美

Google 的未來學家 Ray Kurzweil 預測人工智能將在 2045 年達到「奇點」(Singularity Point)

如果這種情況發生，人工智能系統可能會在無需人類干預的情況下進行遞歸自我改進，從而導致智能和能力的指數級增長

如果失去控制，這是否會造成生存風險，即對人類生存構成威脅



ASIMOV'S THREE LAWS OF ROBOTICS 三大機器人法則

01

不得傷害人類（Do no harm）

- 機器人不得傷害人類，或因不作為使人類受到傷害。

02

服從命令（Obey orders）

- 機器人必須服從人類命令，除非該命令與第一法則相抵觸。

03

自我保護（Self-preservation）

- 機器人必須保護自己，前提是不違背第一與第二法則。

這些法則足夠保障人類嗎？

超級人工智能（ARTIFICIAL SUPERINTELLIGENCE）的風險

人類水平人工智能

具備認知、情感和道德智慧？

人類無法控制

生存風險

應該遵循哪種道德框架

自然律、義務論、功利主義？

大規模失業

全民基本收入？

自主武器與

國際權力鬥爭

機器權利

存在權、自主權、思想自由權、法律保護權

NICK BOSTROM'S PAPERCLIP MAXIMIZER
「萬字夾製造機」思想實驗

- 一個超級人工智能被編程成只追求一個目標 — 生產萬字夾
- 這個超級人工智能會不斷學習與改進，並全身心投入於達成這個單一目標
- 它變得越來越高效，最終壟斷所有可用資源
- 包括那些對人類生存至關重要的資源，只為了最大化萬字夾的生產，將世界變成無盡的萬字夾空間

「我們的未來，很大程度上取決於人工智能一開始怎樣被設計與控制。
這件事我們只有一次機會去做好。第一個超級人工智能，可能會是唯一一個——除非我們能夠掌握其行為。這是我們這個時代最重要的任務。」

– Nick Bostram

人工智能教父 GEOFFREY HINTON – 未來將需要全民基本收入 UNIVERSAL BASIC INCOME

https://www.youtube.com/watch?v=em5sglAMpeo

紅十字國際委員會（ICRC）就自主武器的立場

機器應該做出生死決策嗎？

“Machines will never be able to bring a genuine humanity to their interactions, no matter how good they get at faking it.”

https://www.youtube.com/watch?v=8GwBTFRFlzA

類人AI

- 有「生命」
- 有「意識」
- 能夠「感覺痛苦」
- 能夠「有欲望」
- 能夠「理解因果關係」
- 能夠「有意圖地產生某些結果」
- 是否需要「道德地位」？

機器權利

Human-like AI 是否應該像人類一樣受到尊重，讓它們能追求美好的生活？

生存風險
EXISTENTIAL RISKS

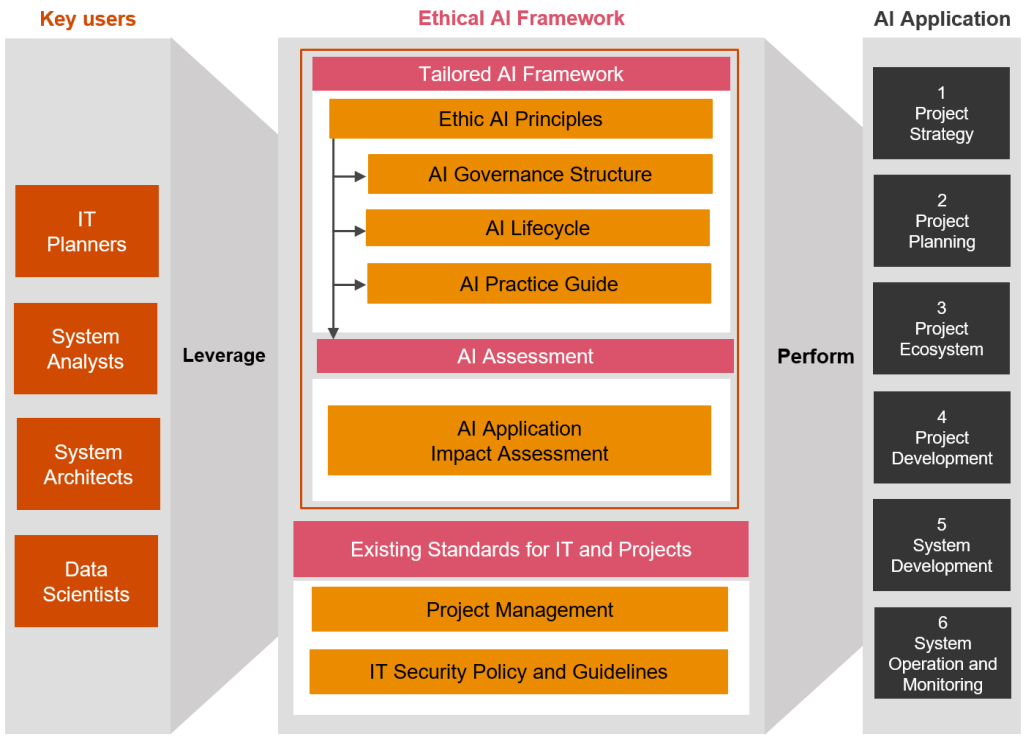
10% - 20% 的機率，人工智能在未來三十年內導致人類滅絕
(Geoffrey Hinton, 2024年12月)

https://www.cbc.ca/player/play/video/9.6602767

STUART RUSSELL'S VALUE ALIGNMENT 三大價值對齊原則

- 機器的唯一目標是實現人類的價值
  - 取代 Asimov 的「服從命令」想法，改為更深層的「價值一致性」
- 機器一開始並不確定人類的價值是什麼
  - 使 AI 願意接受修改及人類的干預/指導
- 機器應透過觀察人類行為來學習人類的價值
  - AI 並非預設知道人類想要什麼，而是須逐步學習人類的真實偏好與價值觀

香港人工智能道德框架



- 透明度與可解釋性
- 可靠性、穩健性與系統保安
- 公平性、多元與共融
- 人類監督
- 合法與合規
- 資料私隱保障
- 安全性
- 問責制
- 合乎公共利益
- 開放合作
- 可持續發展與公義轉型

香港生成式人工智能技術及應用指引（2025年4月發布）

持份者	責任
技術開發者	道德規範的模型開發 技術保障措施的落實 持續監督與監察 數據權利管理
服務提供者	內容治理 私隱保護 問責措施 用戶數據處理
服務使用者	道德使用 意識提升與自我控制 社群保護 內容核實

願基督信徒、不同宗教的追隨者，以及所有善心人士，在和諧中共同努力，迎接數碼革命帶來的各種機會，並面對所帶來的挑戰，讓我們的後代能有一個更精誠團結、更富於正義與和平的世界。

（教宗方濟各2024世界和平日文告）


QESTIONS?