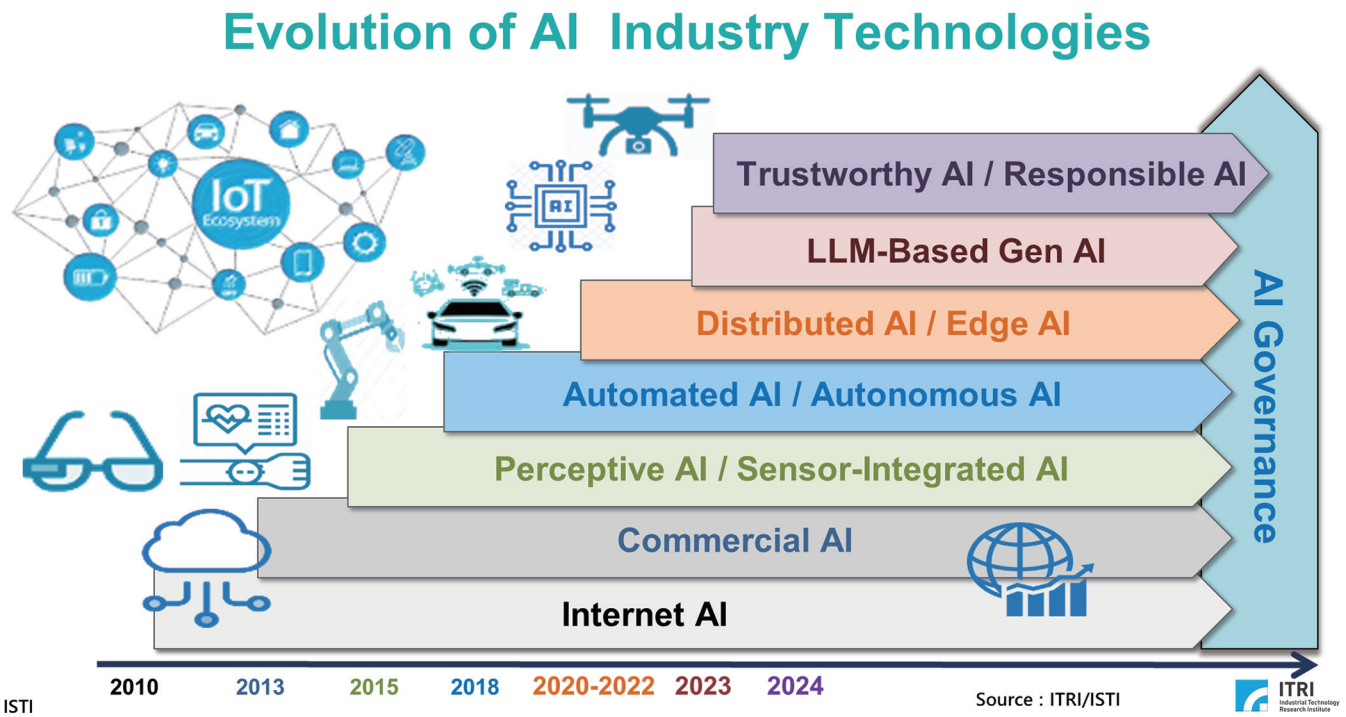
當前人工智慧之技術挑戰與未來方向
Benefits and Risks of Artificial Intelligence

工研院產科國際所研究總監 陳右怡

一、前言：從人工智慧產業技術發展脈絡談起

根據全球人工智慧產業技術發展脈絡，人工智慧技術眞正開始應用到產業市場上起於2010年起，以網際網路應用技術爲主的「Internet AI」，例如搜尋、網路廣告、電子商務、社群媒體、影音內容與遊戲等。然至2014年人工智慧擴大至商業應用場域的「Commercial AI」，此時以銀行、保險、教育、醫療、藥物、物流、供應鏈爲常見導入之應用，而2016年AI導入至物聯網裝置上，以發展AIoT解決方案爲導向的「Perceptive AI / Sensor-Integrated



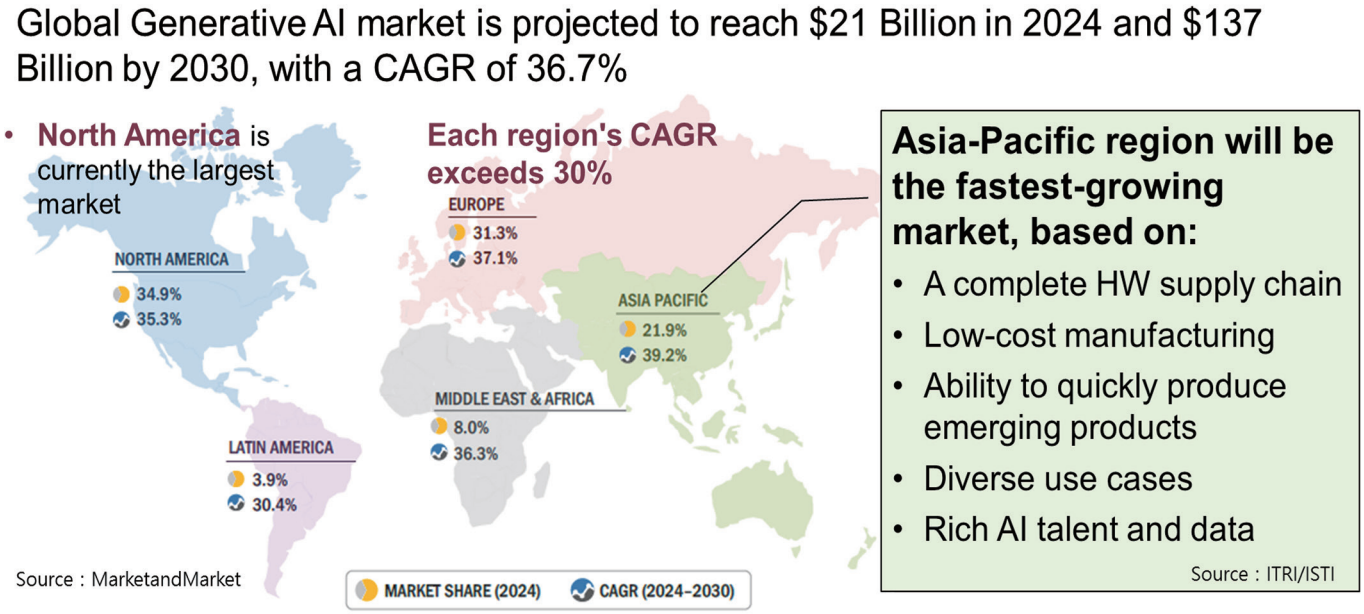
資料來源：工研院產科國際所

AI」，如安全監控、科技零售、智慧居家、智慧城市等；接著2018年開始各種產業開始運用AIoT達到全面自動化之「Automated AI / Autonomous AI」，如智慧製造、智慧倉儲、智慧農業、智慧交通、自駕車、無人機、機器人等應用。從2020到2022近三年新冠疫情、氣候變遷、國際政治經濟及戰爭的動盪下，人工智慧的發展未曾停歇反而加速，正持續進化到「Distributed AI / Edge AI」避免因天災人禍讓企業營運中斷，因此產生分散式運作的架構，達到即時、可靠、穩定、安全之AI運算、處理與分析，在邊緣端完成所有任務，以強化企業營運韌性爲目標；至2022年末Open AI釋出ChatGPT，將人工智慧帶入另一個里程碑，激發「LLM-Based Gen AI」應用熱潮，帶來各種深僞(Deepfake)技術應用的亂象，促使「Trustworthy AI / Responsible AI」的興起，當前的人工智慧技術的技術挑戰在於如何確保 AI從數據、演算法、系統與商業模式等，皆具可解釋性、透明性、可追溯性、安全性。因此，近10年人工智慧應用於不同的產業領域發展快速，「AI Governance」需要考量人工智慧在不同產業領域的應用現況與方向。

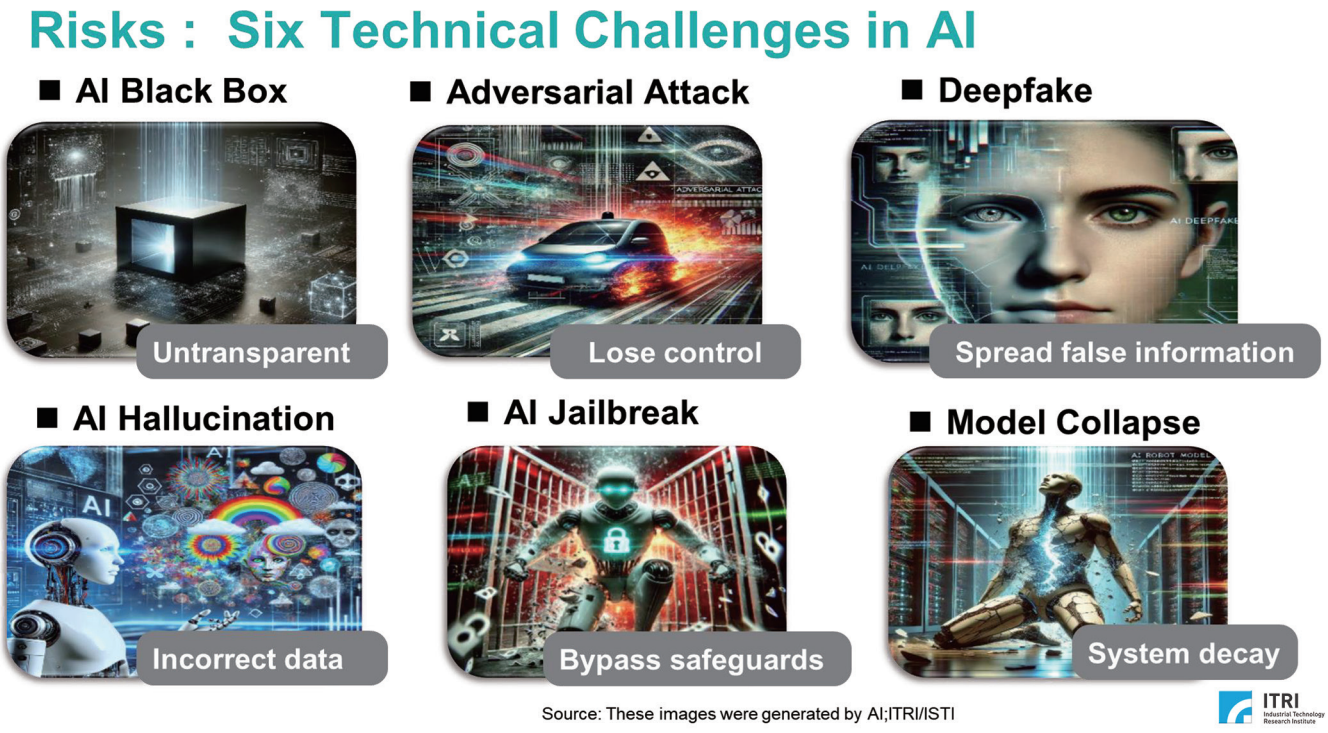
二、人工智慧市場機會分析：亞太地區具備高度成長潛力

特別是2022-2024這三年生成式人工智慧發展速度超乎很多專家的預測，全球生成式人工智慧市場預計到2024年將達到210億美元，到2030年將達到1370億美元，複合年成長率爲36.7%。各區域市場複合年成長率超過30%，其中亞太地區預計將成爲生成式人工智慧成長最快速的市場，此乃基於亞太國家具備有全面性的硬體供應鏈、低成本製造、快速生產新興產品的能力、多樣化的產業用例、豐富的人工智慧人才和數據等優勢。

在人工智慧技術發展推力之下，根據Frost & Sullivan調查，全球近90%的公司將導入人工智慧技術視爲首要任務，而79%的企業已開始實施人工智慧技術，然而，企業之間採用人工智慧技術正處於不同階段—包括從概念驗證到同時部署多個人工智慧用例。儘管採用率不斷提高，但只有20%的企業宣稱達到在企業內外部採用AI已達到無處不在的階段。



資料來源：MarketsandMarkets；工研院產科國際所



資料來源：工研院產科國際所

三、當前人工智慧技術發展問題：各界需正視的六個技術挑戰

當前人工智慧技術仍在快速進化當中，因此當前各界在使用、導入或開發人工智慧技術、產品或服務之際，也需要正視以下這些技術問題或挑戰，其可能造成個人、組織、產業、社會、國家等各種不同層次的風險及負面影響之外，同時，從逆向思維的角度來看，這些技術問題或挑戰也是人工智慧技術創新突破及新創的機會點。如下圖所示，歸納分析如下：

1.AI黑盒子(AI Black Box)

這是人工智慧特點或先天限制。AI模型的運作依賴於大規模數據，特別是深度神經網路(Deep Neural Networks, DNN)，包含大量參數和多層神經網絡，結構複雜難以直觀理解，通常AI採用非線性函數來捕捉數據間的關聯，從中提取的模式往往不符合人類的直覺認知，人類很難拆解AI決策過程，其缺乏透明性與可解釋性。目前各界積極討論並發展可解釋AI(Explainable AI, XAI)技術解決方案，以及可信任AI(Trustworthy AI)的技術規範等，期望各方在使用或開發人工智慧之時，能追溯並釐清各方對AI所應負責的責任，發展可負責任的AI(Responsible AI)，達到人工智慧的可追溯性、可信賴、公平性、穩健度等指標。

2.對抗攻擊(Adversarial Attacks)

這屬於在人工智慧領域中AI模型或系統之安全性問題。透過輸入錯誤的數據或資料，例如圖像、語音或文本等，添加人類難以察覺的微小干擾，欺騙AI系統或攻擊深度學習模型，讓AI系統產生錯誤的判斷或預測結果，導致其失去判斷力和控制能力。這類型的攻擊對於醫療診斷、金融決策和自動駕駛等高風險應用的安全性構成直接威脅。

3.深度偽造(Deepfakes)

此基於生成式人工智慧發展下的產物。"Deepfakes"是由"Deep Learning"和"Fake"合起來的英文字詞，其主要基於生成對抗網路(Generative Adversarial Networks, GANs)等深度學習的模型，合成高度逼真的影像、語音或影片，模仿現實中的人

物或場景，製作出以假亂眞的內容，如複製人臉、聲音、場景、假新聞等假資訊，僞造眞人的身份或聲音，用於金融詐騙、混淆視聽、侵犯隱私等其他犯罪行爲。

4.AI幻覺(AI Hallucination)

多半指的是生成式人工智慧系統或模型，由於訓練數據的局限性、對上下文理解不足、對輸入的指令了解偏誤等問題，因而促使AI生成虛假的、不眞實的或不相關的內容，從而削弱信任並影響決策。AI幻覺所生成的內容以多樣化表現形式下，應再結合人類判斷來進行交叉驗證，以避免受到虛假資訊的誤導。目前技術解決方式是可透過提升訓練AI數據的品質、提升輸入內容的眞實性、優化AI模型架構等方式來減少AI幻覺的現象。

5.AI越獄(AI Jailbreaks)

這是生成式人工智慧使用過程中所帶來的技術挑戰。是指透過特定方法或手段，繞過AI系統內建的限制或安全機制，使得AI產生不符合原先所設計的輸出內容。最常見的AI越獄方式是一般使用者透過提示詞(Prompt)、多層提示(Chain of Prompts)、角色扮演(Role-Playing)、利用模型漏洞(Model Exploitation)等設計出間接提問、語意模糊化、假設方式等，技巧性地引導AI生成如侵犯個人或企業的隱私資料、暴力、歧視、非法活動、違反倫理等相關資料。

6.模型崩潰(Model Collapse)

也稱爲「模式退化(Mode Collapse)」，此屬於生成式人工智慧領域的一種現象。意指AI模型在訓練或應用過程中，由於過度依賴AI自身所生成數據或資料進行再訓練，除了訓練不足(Underfitting)或訓練數據偏差的問題之外，也會因資料迴圈(Feedback Loop)、過度優化(Over-Optimization) 而導致AI模型性能逐漸下降，因而重複生成同質化的內容或輸出品質低劣。崩潰的AI模型將帶來錯誤的資訊傳播和決策、無法提供具有商業競爭力的內容。目前比較常見的解決途徑包括將原始數據結合生成數據進行混合訓練、動態提高眞實數據的學習率、數據多樣化、持續監測AI模型的生成品質等，確保AI模型的穩定度。

四、AI 人機協作發展方向：以 Human-in-the-Loop 爲基本架構

數位時代下各式各樣的智慧機器早已是人類生活中不可或缺的一部分，時常要與各種數位化的裝置、設備與服務等進行人機互動與協作。AI時代下，AI在人機協作的角色有三種層次，說明如下：

1.Human-in-the-Loop：AI的角色是輔助人類決策。人類積極介入並監督AI，保留對AI完全的控制權，AI只需提供建議。若無人類積極作爲，AI無法決策。

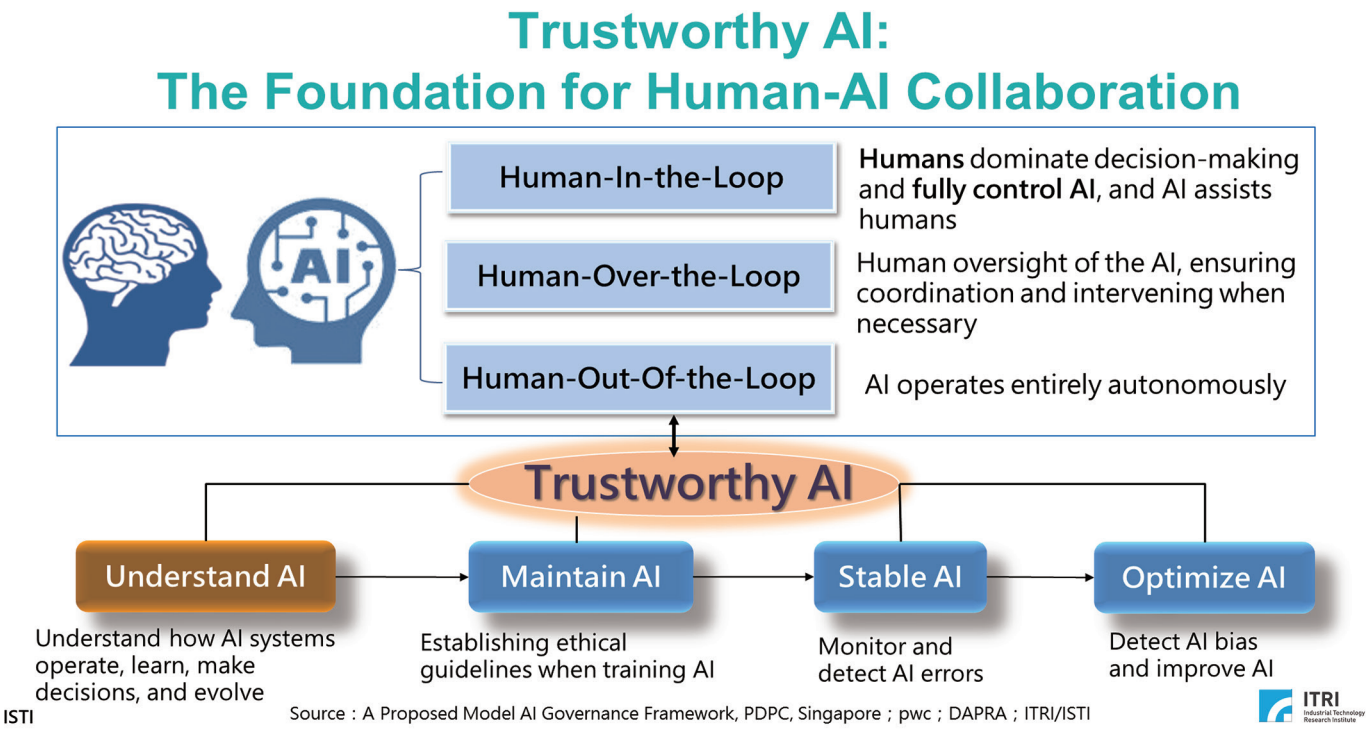
2.Human-Over-the-Loop：人類可調整AI執行決策。允許人類在AI執行演算法等任務時，也能調整參數等。

3.Human-Out-Of-the-Loop：由AI完全決策。AI執行決策時，完全沒有人類監督，由AI完全控制，人類不能介入AI決策的選項。

在AI角色層次不同的狀況下，各方對人工智慧的看法將左右整個AI系統或模型的發展。因爲在AI系統的運作過程中，不論是資料的品質、演算法的設計，或是人爲操控等，都可能造成訓練或生成的結果有所偏頗，甚至資訊安全、隱私保護、風險識別等問題，就必須藉由「可信任AI」的技術與管理機制。目前全球產官學研已開始對開發AI、應用AI的責任進行原則規範。要求進行AI風險評估以符合各種信任需求，這將是未來全球各國發展人工智慧技術、產品或服務的機會與挑戰。因此，本文再展開並歸納出可信任AI」四個發展途徑：

1.理解AI：可追溯因果以及制定規範。理解AI系統如何運作、學習、決策、進化，同時掌握因AI引發的因果關係及所有參與者的責任歸屬。

2.維護AI：訓練AI時建立道德指導原則：可預測性、可重複性。能追溯大數據假設盲點及模型在某



資料來源：A Proposed Model AI Governance Framework, PDPC, Singapore；pwc；DAPRA；工研院產科國際所

些條件下會失效或失敗，以採取適當的系統措施。

3.穩定AI：監控並偵測AI錯誤。能採取糾正措施或關閉AI系統，並能辨識有危害系統或影響安全之不良設計、駭客攻擊、侵犯隱私。

4.優化AI：檢測AI偏差並改善AI。找出ML模型中的缺陷與大數據中的偏差，能驗證AI預測、改進模型、決策過程、獲得新見解。

五、結論與建議

1.每個人對人工智慧的看法都不同，但共同理解AI至關重要

當前人工智慧的發展已全面席捲人類的日常生活、知識學習、工作與社交等層面，運用AI工具或能與AI共同協作將是必備技能。現在嬰幼兒、學齡兒童是人工智慧原住民，將與AI共同成長、學習與協作。可以預見的是，從個人到社會將擁抱AI工具，並從中了解AI的能與不能，了解AI才能善用AI。建議台灣AI產品或服務能根據不同使用族群的不同使用情境下設計，發展以人類使用者爲中心的AI人機協作介面或流程。

2.「產業AI化」、「AI產業化」到「AI平民化」

不論是協助各行各業導入AI創新轉型的「產業AI化」，或者創造出AI新產業的「AI產業化」，生成式AI加速「AI平民化」時代的來臨，也就是不論是國際大廠還是中小企業皆可從人工智慧中受惠如獲得商機或加速營運效率等、而一般使用者也能使用到人工智慧以追求工作或生活品質，因此走向AI everywhere！

3.進一步思考AI跨域前瞻技術，以提早深耕布局AI特定領域技術

AI激發各種技術創新及融合，因此除了關注當前備受產業注目的AI熱門技術，例如AI半導體、邊緣AI、通用AI、產業專用模型、多模態模型、AI代理人等發展，建議台灣可進一步再思考發展跨域前瞻AI技術領域，例如追求低功耗且高效能、易於模組化、模仿人類大腦運作機制之演算法及運算設計、AI新腦機協作介面、以及能多方支援Multi-Modal AI模型等。