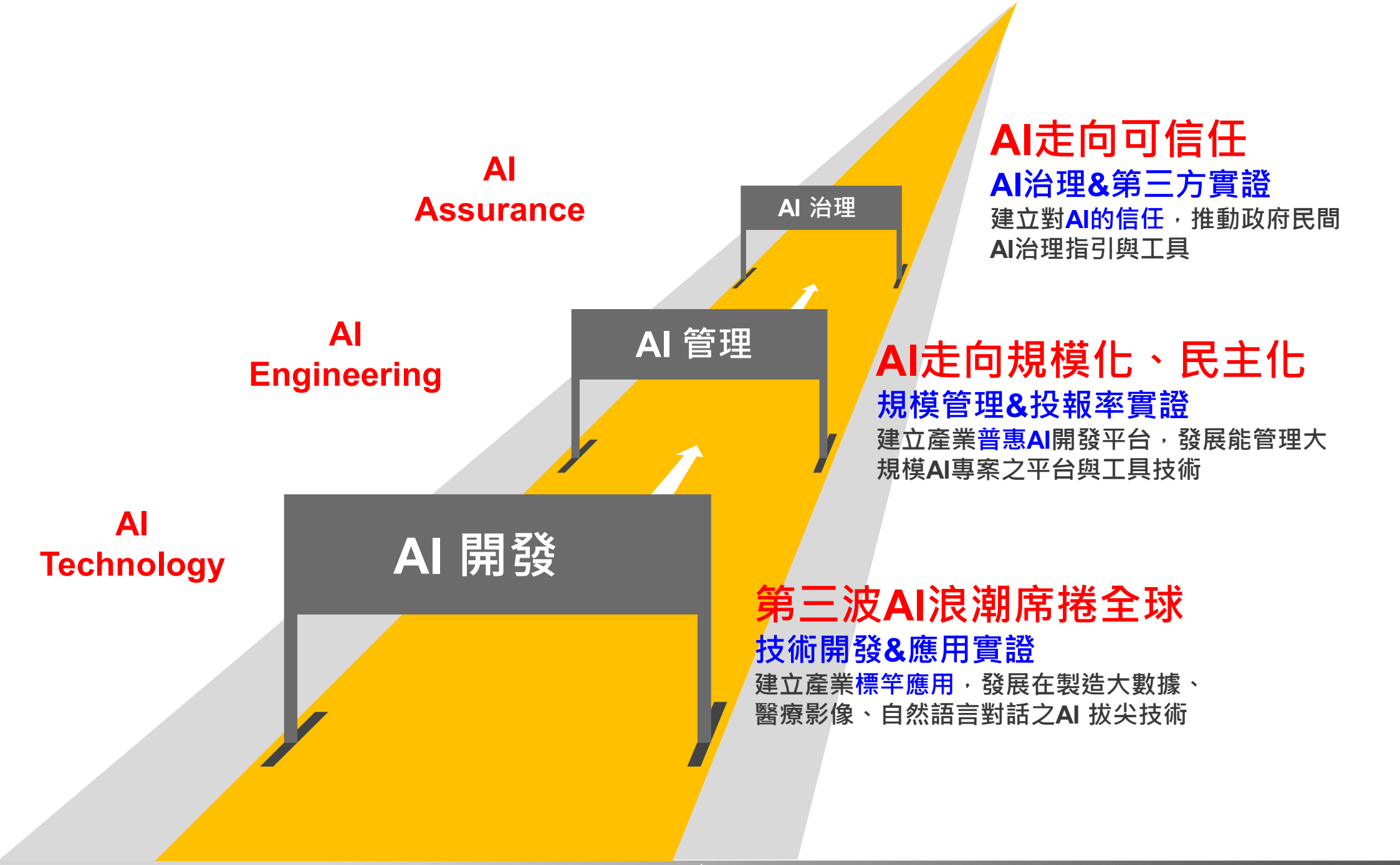
AI發展的下一步：可信任AI

何文楨 資深技術總監
軟體技術研究院

2022.07.22

wenjen@iii.org.tw
www.iii.org.tw

AI發展下一步？



AI走向可信任

AI治理&第三方實證

建立對AI的信任，推動政府民間AI治理指引與工具

AI走向規模化、民主化

規模管理&投報率實證

建立產業普惠AI開發平台，發展能管理大規模AI專案之平台與工具技術

第三波AI浪潮席捲全球

技術開發&應用實證

建立產業標竿應用，發展在製造大數據、醫療影像、自然語言對話之AI 拔尖技術

可信任AI同義詞

資料來源: 工研院產科國際所

可信任AI原則

全球36份公開文件

角色定位

- 公民社會(如UNI Global Union)
- 政府
- 政府機構(如OECD、G20)
- 利益團體(如產學研、產業聯盟、IEEE等)
- 私人企業(如IBM、Google)

排名	AI原則	%
1	Fairness and Non-discrimination	100%
2	Privacy	97%
2	Accountability	97%
3	Transparency and Explainability	94%
4	Safety and Security	81%
5	Professional Responsibility	78%
6	Human Control of Technology	69%
6	Promotion of Human Values	69%

資料來源: Harvard Uni., 工研院產科國際所整理

可信任AI原則

1. **公平性與非歧視性 (Fairness and Non-discrimination)**
   - 說明: 在編寫AI演算與決策時，能避免因為不公平或是歧視等偏見導致運算結果的偏差
   - 舉例: 因為男女性別、黑人白人膚色的不同，對於保險費率的估算造成偏差
2. **透明性與可追溯性 (Transparency and Traceability)**
   - 說明: 對於AI技術所生成的決策過程採行的AI系統、軟體及演算法等技術與應用，要能建立紀錄保存制度。若發生爭議時，能有依據使受到 AI 技術決策影響之利害關係人得以釐清或事後救濟。
   - 舉例: AI輔助做交通違規的裁決判斷時，能公開AI決策判斷的演算方式
3. **個人隱私與數據治理 (Privacy and Data Governance)**
   - 說明: AI研發與應用時所使用的資料，在蒐集處理利用的過程都須符合相關的法性規範，以確保資料當事人的權益
   - 舉例: 對於醫療影像的AI運算，採用的病理資料能將個人身分資料去識別化後再使用
4. **安全性 (Safety)**
   - 說明: 考量網路與資訊安全、風險控管與監測等因素，建構合理善意且安全可靠的AI技術運作的環境
   - 舉例: 保全監視器蒐集到的影像資料，能確保傳輸與使用過程中不會遭受到不當攻擊與外洩

可信任AI原則

5. **問責與溝通 (Accountability and Communication)**
   - 說明: 為維護社會與利害關係人的權益，對於AI發展與應用應有相關機構，敦促與監理AI產業發展
   - 舉例: 政府新設立的數位發展部，將AI產業發展納入國家數位治理的重點產業之一
6. **自主權與控制權 (Autonomy and Control)**
   - 說明: 對於AI技術所生成的決策，AI技術提供的是決策的建議，最終的決策選擇權還是人類
   - 舉例: google map參考交通工具、路徑距離、運輸時刻表、行進時間、收費與否等因素提供的交通路線建議，最終採用何種方式抵達目的地的決定權還是在使用者本身，而非系統
7. **可解釋性 (Explainability)**
   - 說明: 對於AI技術所生成的決策，要能提供發展脈絡的相關資料，以利對於利害關係人進行事後的說明、展現與解釋
   - 舉例: 車輛使用先進駕駛輔助系統(ADAS)時，對於行車決策的判斷能提供完整的解釋脈絡
8. **共榮共利 (Common Good and Well-being)**
   - 說明: AI應該要能協助人類保障身心健康，創建社會更高利益、提升總體環境發展
   - 舉例: AI可以減輕蒐集大量資料與精密運算的時間，如資料文獻蒐集分析、數據分析演算，人類可以花更多時間在發想與創新等對未來更有啟發性的工作

可信任AI的建置

- 根據不同AI信任議題發展技術工具

階段	模型建置前	模型建置中	模型建置後
目標	識別並避免使用有偏見的資料，建立資料的可信任度	發展更多結構化,因果模型等技術，建立具備可解釋性的AI模型	為沒有提供解釋的AI提供解釋，透過模型監控建立可信任依據
原則	• 公平性 (Fairness) • 隱私權 (Privacy) • 資料偏誤 (Bias)	• 透明性 (Transparency) • 解釋性 (Explainability) • 測試與評估 (Testing & Evaluation) • 驗證與確認 (Verification & Validation)	• 穩健性 (Robustness) • 可靠性 (Reliability) • 安全性 (Safety & Security) • 可問責 (Accountability)

研發開源工具套件

IBM

python GitHub

- AI Explainability 360：提供的演算法涵蓋了ML模型的直接解釋性和間接解釋性指標，可應用於金融、人資、教育、醫療保健等領域
- AI Fairness 360：提供度量標準以檢查數據集和機器學習模型中是否存在不必要的偏差，並包含減輕偏差的九種算法，可應用於信用評分、醫療費用預測、人臉圖像分類
- Adersrial Robustness Toolbox：支援開發人員和研究人員加強深層神經網絡(DNN)受到對抗性攻擊安全性的工具庫

資料來源 MIC 整理, 2021

可信任AI開源軟體

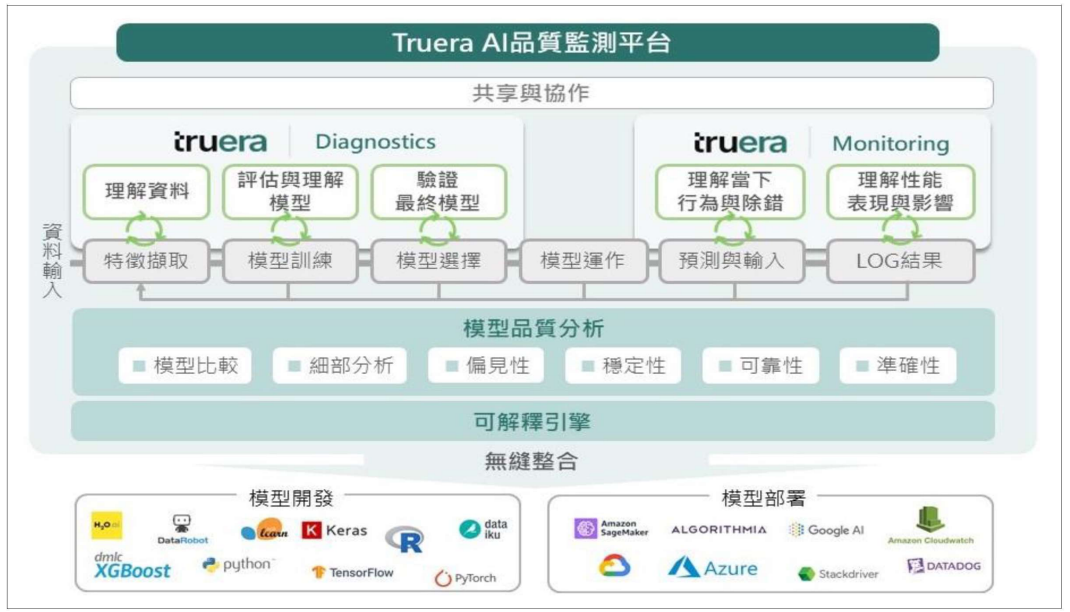
Objective	Tool	Description
Fairness	AT&T software System to Integrate Fairness Transparently (SIFT)	Software system to integrate mechanised and human-in-the-loop components in bias detection, mitigation, and documentation of projects at various stages of the machine learning lifecycle.
	Microsoft Fairlearn	Open-source toolkit to assess and improve the fairness of machine learning models. Contains an interactive visualisation dashboard and bias mitigation algorithms to help navigate trade-offs between fairness and model performance.
	LinkedIn Fairness Toolkit (LiFT)	Open-source toolkit to enable measurement of fairness according to a multitude of fairness definitions in large-scale machine learning workflows.
	Google What-If Tool	Open-source software tool to visually inspect and explore machine learning model performance and data across multiple hypothetical situations, with minimal coding required.
	IBM AI Fairness 360	Open-source toolkit to help detect and mitigate unwanted bias in machine learning models and datasets. Provides approximately 70 metrics to test for biases, and 10 algorithms to mitigate bias in datasets and models.
Transparency	IEEE Standard for Transparency of Autonomous Systems	Technical standard to describe measurable and testable levels of transparency, so that autonomous systems can be assessed and levels of compliance determined.
	Google Model Card Toolkit	Documentation framework for sharing the essential facts of a machine learning model in a structured, accessible way, providing an overview of what the model is intended to do, how it was architected, trained, and its limitations.
Explainability	Google Cloud Explainable AI service	Software to help developers get explanations on the outcomes of their models. Can be applied to the AI models trained on tabular, image, and text data. Not open source.
	IBM AI Explainability 360 Toolkit	Open-source toolkit of algorithms, code, guides, tutorials, and demos to support the interpretability and explainability of machine learning models.
	Microsoft InterpretML	Open-source toolkit containing machine learning interpretability algorithms to help understand model predictions.
Robustness	IBM Adversarial Robustness 360 Toolkit	Open-source toolkit for machine learning security. It provides tools to evaluate, defend, certify and verify machine learning models and applications against the adversarial threats of evasion, poisoning, extraction, and inference.

Source: Tools for trustworthy AI: *A framework to compare implementation tools for trustworthy AI systems, OECD, 2021*

可信任AI新創

- **Privacy**
  - Owkin
    - 美國新創，聯合學習，去中心化醫療資料集
  - Hazy
    - 英國新創，GAN ，合成資料
  - DataGen
    - 以色列新創，GAN，3D 合成圖像資料
- **Monitoring**
  - Truera
    - 美國新創，MLOps，AI 監控
- **Explainability**
  - Hacarus
    - 日本新創，Sparse Modeling，AOI
  - Kyndi
    - 美國新創，Knowledge Graph，非結構文本分析
  - Darwin AI
    - 英國新創， Neuron Deletion ，AI 模型開發

DataGen 生成不同種族人臉的合成圖像資料



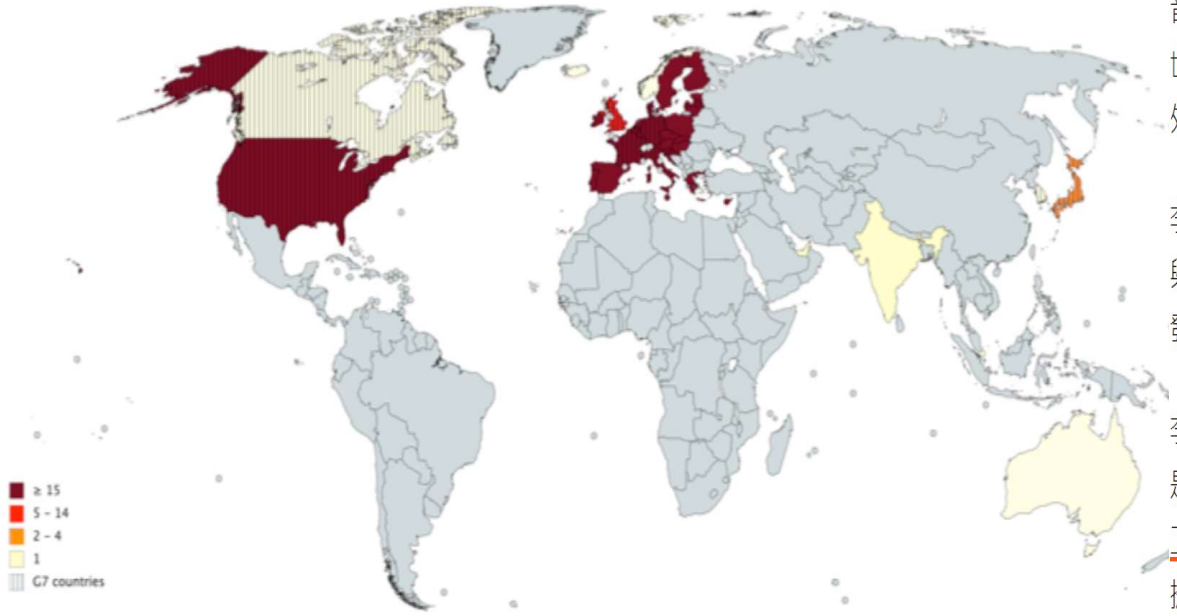
Truera AI品質監測平台

資料來源 MIC 整理, 2021

國際趨勢與影響

- AI正以一種驚人的速度在發展，為AI制定的規範數量越來越多，涵蓋面越來越廣，逐漸從**非強制性的自治**轉變成**具強制力的規範**
- 我國AI相關產業在放眼國際市場的同時，**需積極面對各國制定的規範**，推出符合國際規範之可信任AI，避免受到過大的衝擊與影響

全球AI倫理準則分布總覽(數量與地區)



Source: The global landscape of ai ethics guidelines. Nature Machine Intelligence (2019)

鴻海研究院: **台灣AI產品外銷 須切入各國驗證機制**

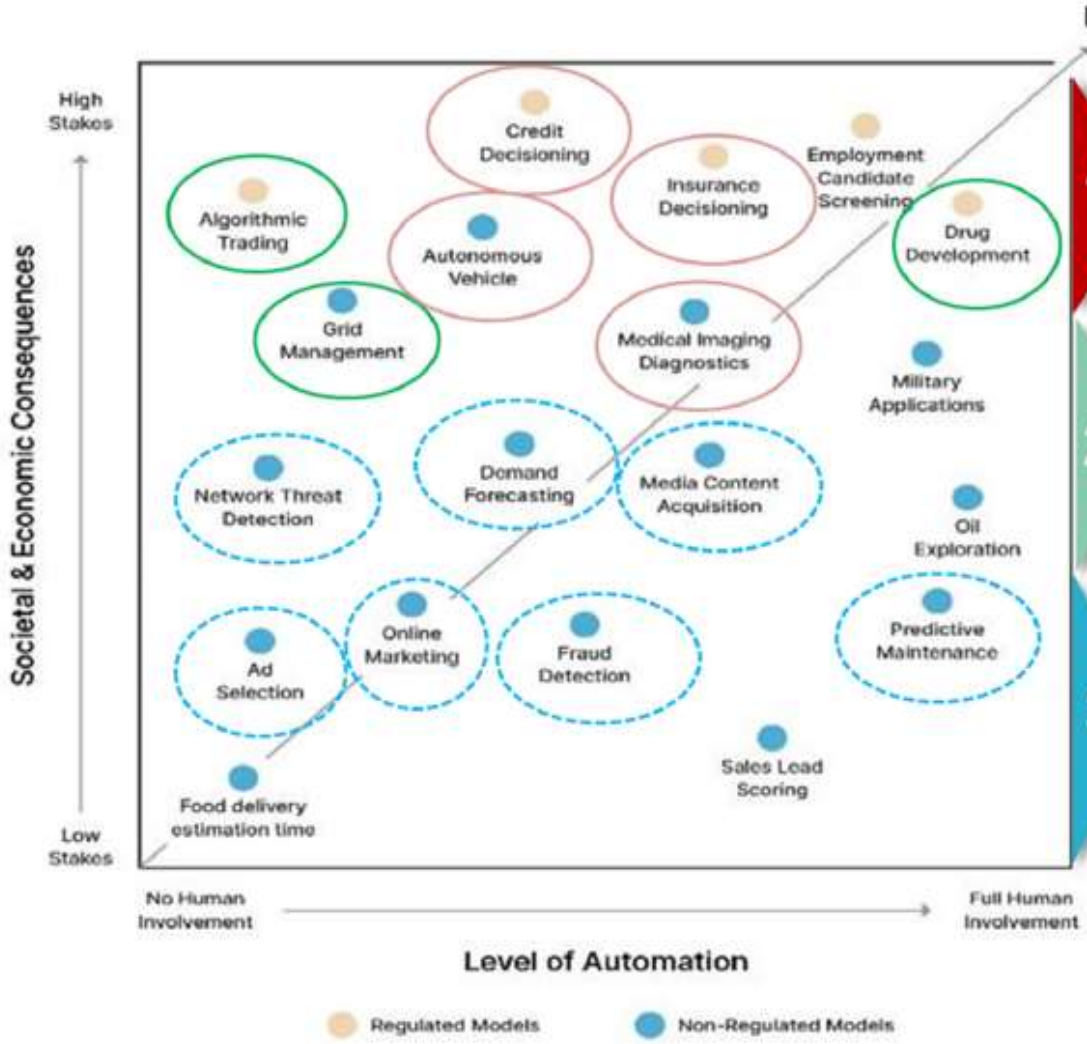
〔記者方韋傑 / 台北報導〕鴻海研究院執行長李維斌今天出席由DIGITIMES、科技報橘、IC之音、人工智慧基金會舉辦的2022 Taiwan AI EXPO，李主講「駛進AI：重新思考AI賦能的未來世界」時指出，現在歐美與澳洲等國皆投入研發人工智慧（AI），台灣要將相關產品銷往海外、打國際戰，必須切入各國驗證機制。

李維斌說，30年前的網路就如同今天的人工智慧，讓大家抱有想像空間，當中存有許多的潛力與機會，有價值、受重視是人工智慧的核心，從以往網路的發展歷程來對照目前人工智慧初期發展階段，要將安全性納入應用考量，忽視這點將付出比發展網路更大的代價。

李維斌形容網路安全如同蟑螂，在人工智慧發展的過程中，會發生在各類應用場域之中，不論是產業人工智慧化、人工智慧產業化，人工智慧的資安議題都相當重要，未來是否能夠信任人工智慧相關系統，會成為應用上的關鍵因素，只談論技術相對單純，但當遇上資安議題將顯著提升人工智慧應用複雜程度。

資料來源: https://ec.ltn.com.tw/article/breakingnews/3914721

可能受影響的產業



1 銀行、保險、醫療、自駕車或受管制產業等

-信用風險評估及決策、保單決策

2 高風險應用

-資產管理、演算法交易、能源投資、藥物開發

3 人類行為決策及商業流程

-預測性維護、詐欺偵測、需求預測、線上廣告、商品促銷、行銷、潛在顧客評分

資料來源: Truera, 工研院產科國際所整理

國際因應作法

國際規範 → **風險定義** → **風險評量** → **風險分級** → **採取措施**

各國政策、法規、制度等，並且學習政府機關、民間企業、人民的因應

參考現行國際規範，定義AI風險領域、緣由、特性等，進而訂定評量準則

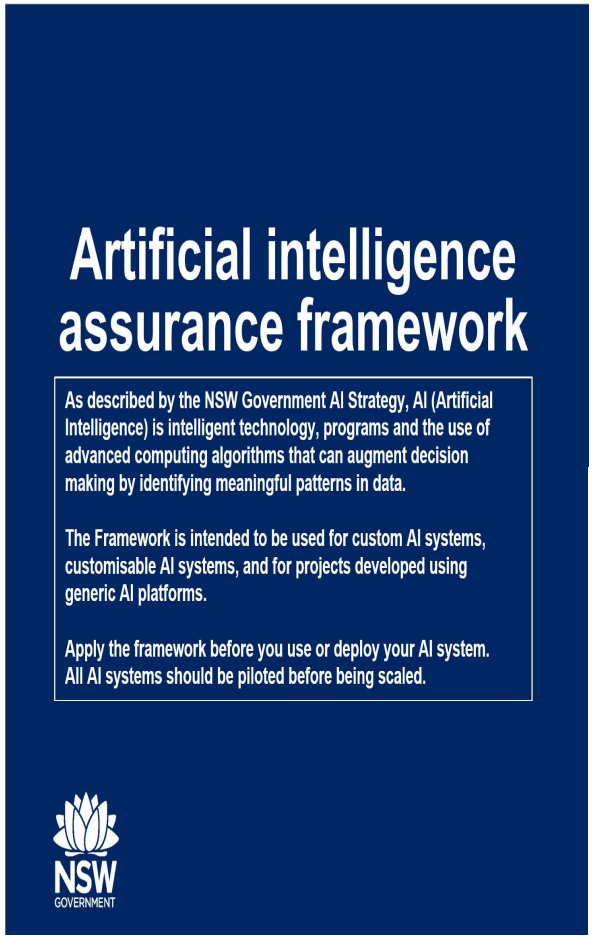
依據準則執行檢測進行AI評量，從規劃、環境建置、評測程序，並提供客觀結果

根據AI評量結果確認在不同領域的風險等級，並且提供相關建議和指南

保障供給端AI產品，需求端安心使用AI產品

國際案例：澳洲

- **澳州 NSW AI Assurance Framework**
  - 幫助政府部門設計、建構和使用AI的產品和解決方案等
  - 採用**風險管理**概念，藉由風險評估，掌握潛在風險，預先採取措施



國際案例：英國

- **英國 ICO AI and Data Protection Toolkit**
  - 一種**風險評估工具**，在AI系統生命週期的不同階段使用

From Principle to Algorithm

評量過程

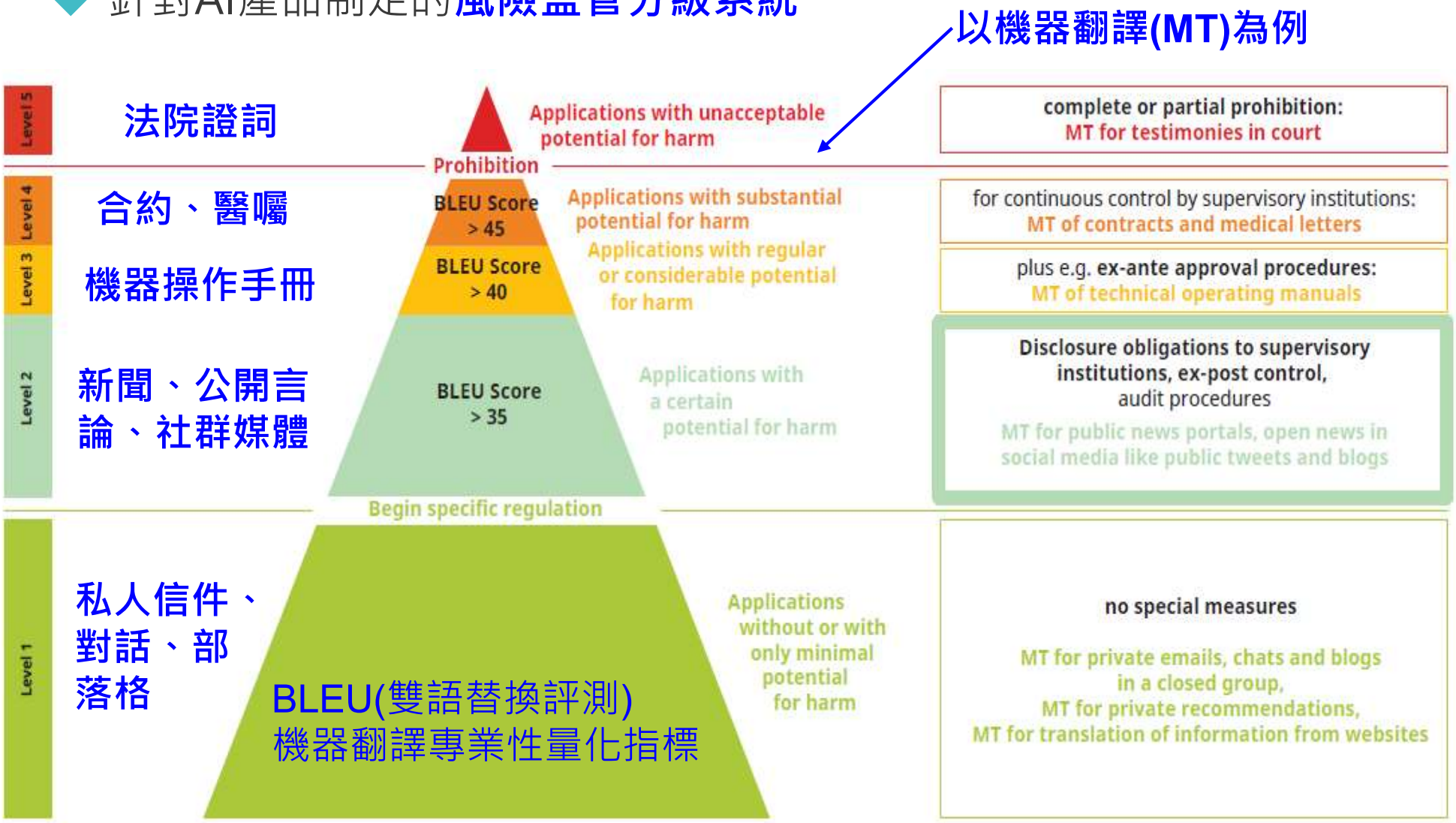
AI生命週期	風險領域	風險描述 / 風險緣由	評量指標 (機率、嚴重性、風險程度)	評量程序
				步驟1: 分析資料集
				步驟2: 分析AI模型架構
				步驟3: 確認模型參數設定…
				步驟4: 檢驗操作架構

Lifecycle Stage	Risk domain area	Risk Statement	How AI can create or exacerbate this risk	Probability (1 = Low, 2 =	Severity (1= low, 2 =	Risk score (probability x	Steps (Technology)
Business Requirements and Design	Fairness (Statistical accuracy, bias and discrimination)	Inaccurate outputs or decisions made by AI systems caused by insufficiently diverse training data, training data that reflects past discrimination, design architecture choices or another reason. This leads to adverse impacts on individuals such as discrimination, financial loss or other significant economic or social disadvantages.	AI systems may use other data points as a proxy for characteristics leading to unlawful discriminatory outcomes that are difficult to detect.	1	1	1	**Step 1**: Map out the purposes for the system, including any decisions that will be made about individuals based on, or influence by, the AI system, as well as the different outcomes and their effects on those individuals. Conduct an initial assessment of potential forms of statistical inaccuracies (including unfair bias and discrimination), which include how you will meet your fairness requirements in relation to discrimination in your context. This should include your mitigation and management strategies. Ensure that risks are drawn from a wide range of stakeholders including policy, user research and design, computer science expertise, and data subjects (or their representatives). Ensure that your assessment is conducted by appropriately skilled personnel (this may require a cross-disciplinary approach, eg data scientists working with legal counsel and review boards). Your assessment should be understood and signed off by an appropriately senior personnel. **Step 2**: Document what the minimum success criteria that is necessary to proceed to the next step of the lifecycle (eg minimum statistical accuracy achieved in the testing phase before proceeding to deployment, or minimum level of fairness based on a specific fairness metric). You should consult with domain experts to inform you which metrics are contextually most appropriate for the model. You should initially focus on outcomes that are immediately experienced by individuals and whether they would reasonably expect the outcomes, or whether adverse effects could be justified. You should also consider the different impacts of false positive and false negative outcomes. **Step 3**: Ensure the team that will be responsible for building the AI system have an awareness of the assessment that has taken place and what their requirements are.

包含10 個符合英國 GDPR 原則的風險領域

國際案例：德國

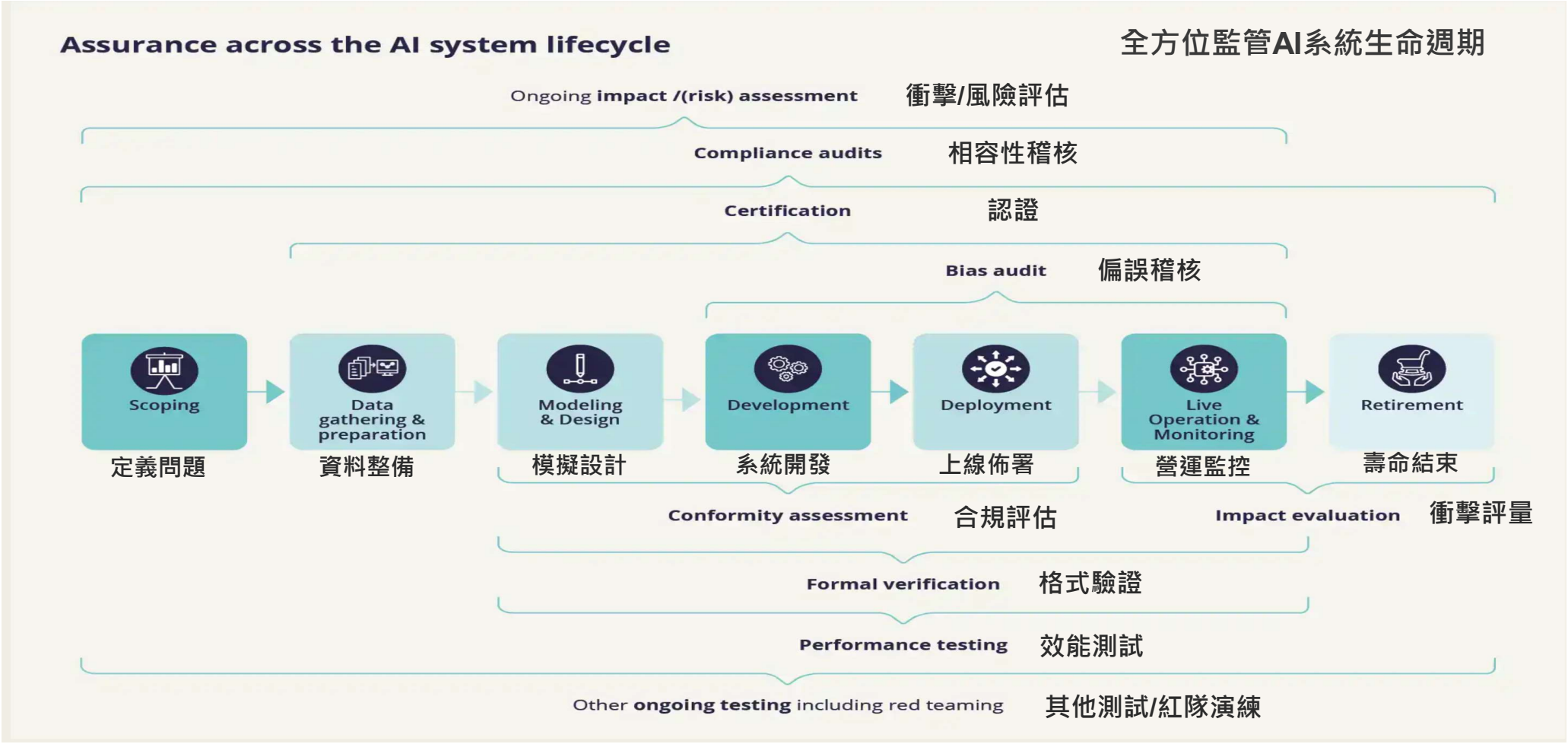
- 德國 DKE AI Risk-adapted Regulatory System
  - 針對AI產品制定的**風險監管分級系統**



Source: German Standardization Roadmap on Artificial Intelligence, Nov., 2020

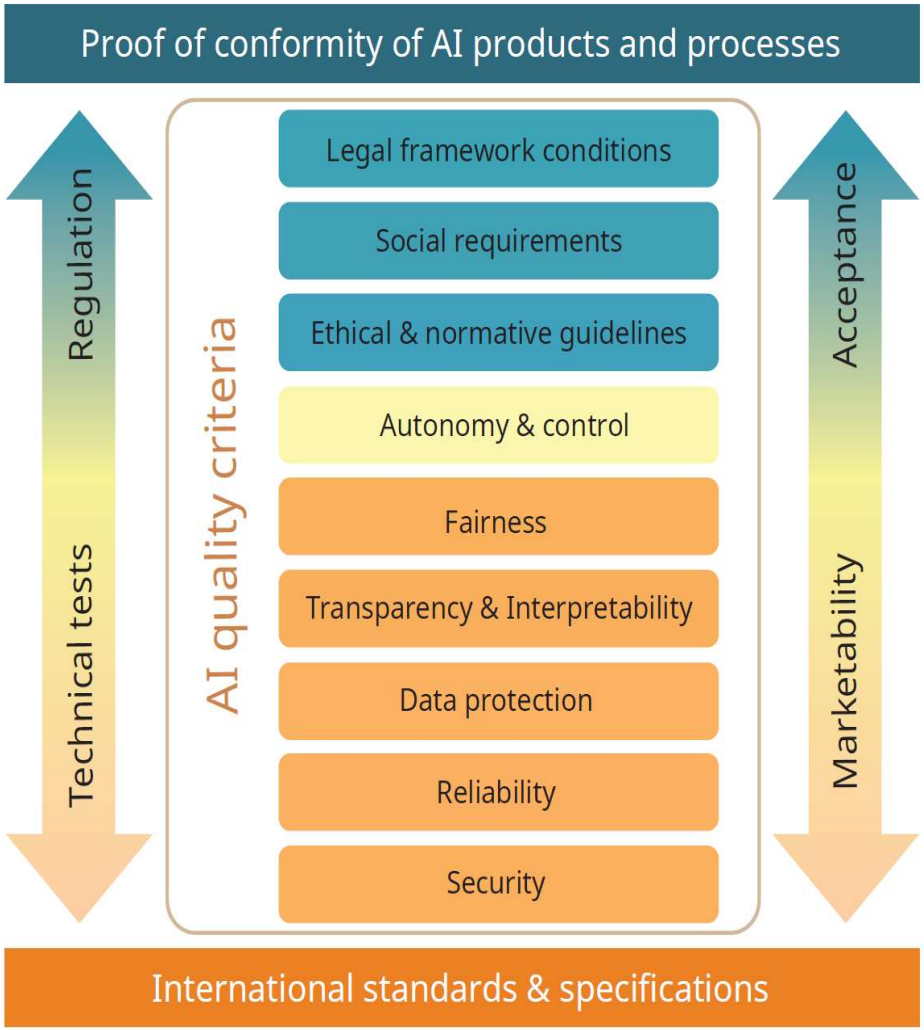
AI→可信任AI

- 光靠團隊的自我認知與自律是不夠的，必須取得團隊外部專業且具公信力的品質憑證，才有可能贏回對AI的信任
- 透過**AI系統生命週期監管**建立信任依據的AI產品/系統

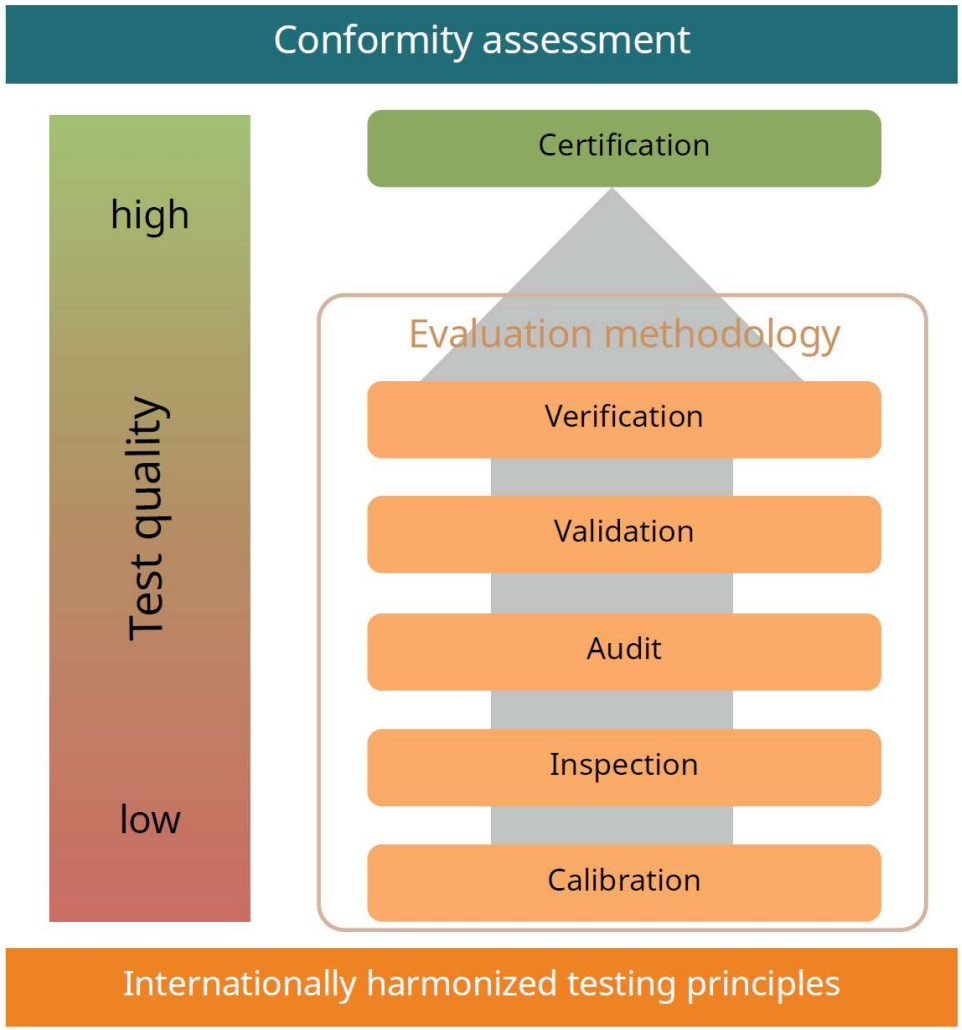


Source: https://cdeiuk.github.io/ai-assurance-guide/

可信任AI合規測試



合規測試-類別項目



合規測試-評估方式

Source: German Standardization Roadmap on Artificial Intelligence

可信任AI的落實

- Provide comprehensive quality assurance for AI solution
  - Data, Model, Process, Performance, Security

1. Assure Data Quality

Test data/feature for correctness.
Completeness and quality.

- 清除任何因數據不足而無法訓練 / 測試的偏差。
- 輔助模型在開發及部署階段有良好的效能。

2. Assure ML Model Quality

Test/ evaluate ML model for it's correctness.
Completeness and quality.
Completeness and quality.

- 確認選擇的演算法適合當前待解問題。
- 確認數據驅動的決策提供正確的結果。

3. Assess ML Process Quality

Access ML Process and frameworks form quality assurance perspective.

- 驗證期流程是否符合業界流程。
- 確保流程是高效率的。

4. Assure ML Performance, Security Quality

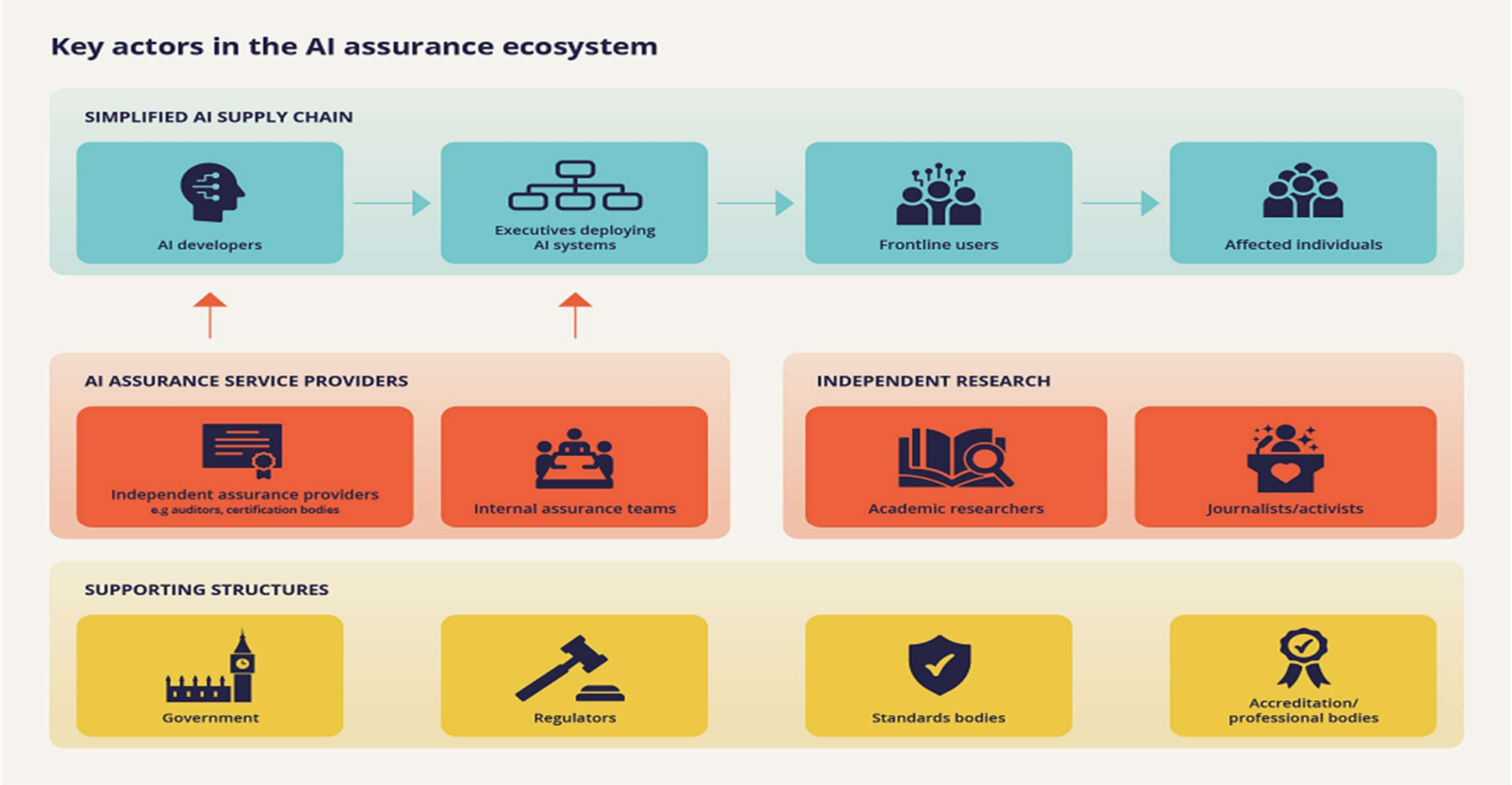
Assure ML solution on non-functional aspects including security and performance.

- 確保模型的訓練和部署都在安全的軟體框架中。
- 確認模型對數據異值及不同質量具適應性。

360° QA for AI

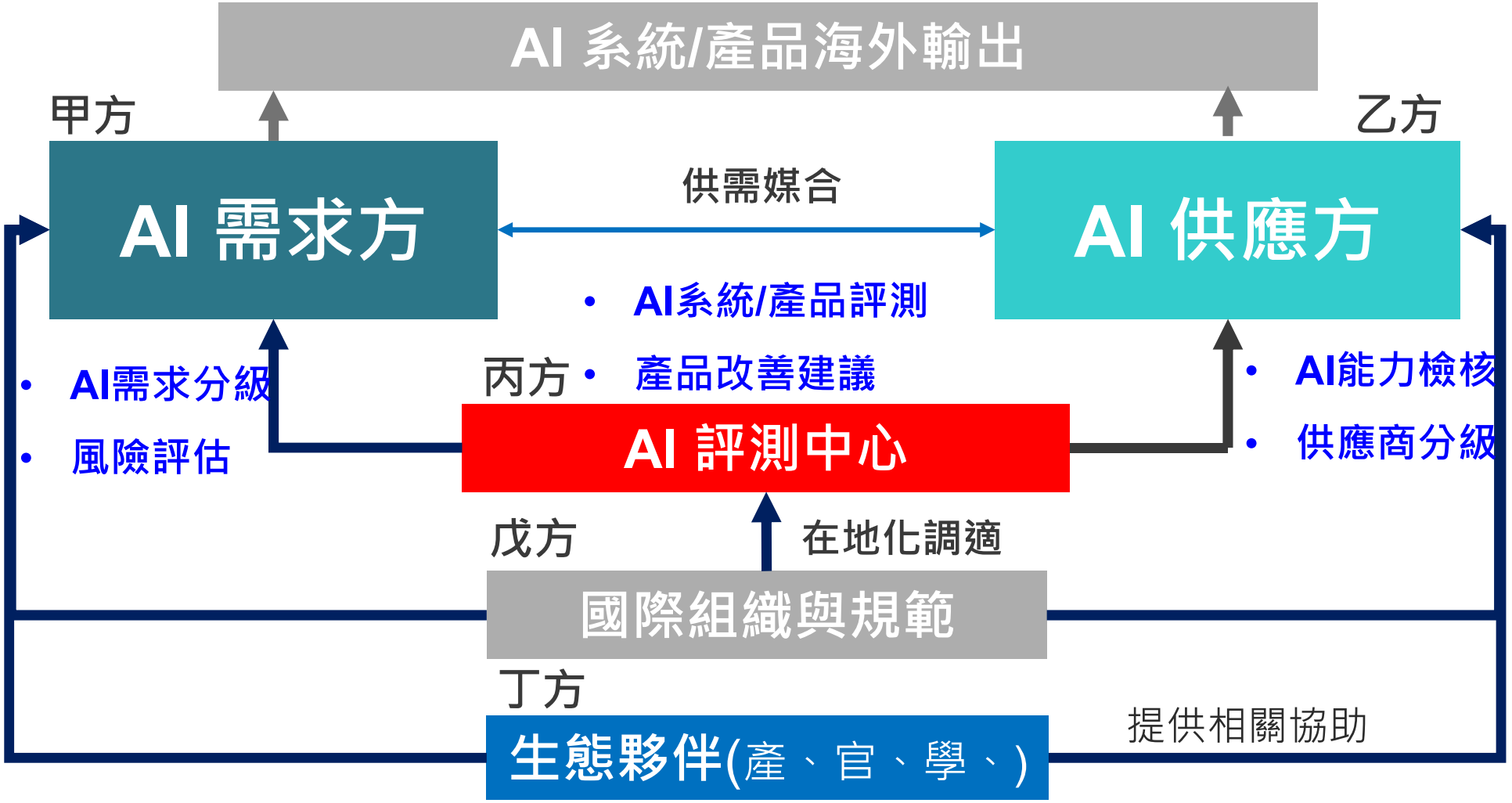
可信任AI生態系

- 生態系的角色
  - AI 產業鏈、可信任AI服務提供商、支持體系、研究機構



Source: The roadmap to an effective AI assurance ecosystem, CDEI, UK, 2021

可信任AI生態系推動架構



結論: 健全可信任AI生態系

- 英、美、德等國的**AI未來發展策略**都將可信任AI當成保持AI領先地位的重要關鍵
  - UK: The roadmap to an effective AI assurance ecosystem
  - US: Advancing trustworthy AI
  - DE: German Standardization Roadmap on Artificial Intelligence
- 目前在建立可信任AI生態系的**努力是分散的**，為了使生態系能有效運作，在兼顧降低風險與社會危害下，又能支持新創創意的發展，需要生態系不同的角色建立共識，一起努力達成
- **健全服務生態系是可信任AI成功關鍵**，强大的制度、工具和服務才能確保可信任IA順利推展
- **可信任AI市場仍處萌芽階段**，對台灣AI產業是很好的發展機會，從賺技術財轉為賺管理財，提早布局，取得關鍵地位