AI 協作治理新篇章

從治理框架到評測驗證的實踐路徑

數位發展部 | 政務次長 侯宜秀

2026 · 05 · 13 · 114年度數位治理研析成果發表會

ISABEL HOU

Deputy Minister, Ministry of Digital Affairs (MODA), Taiwan

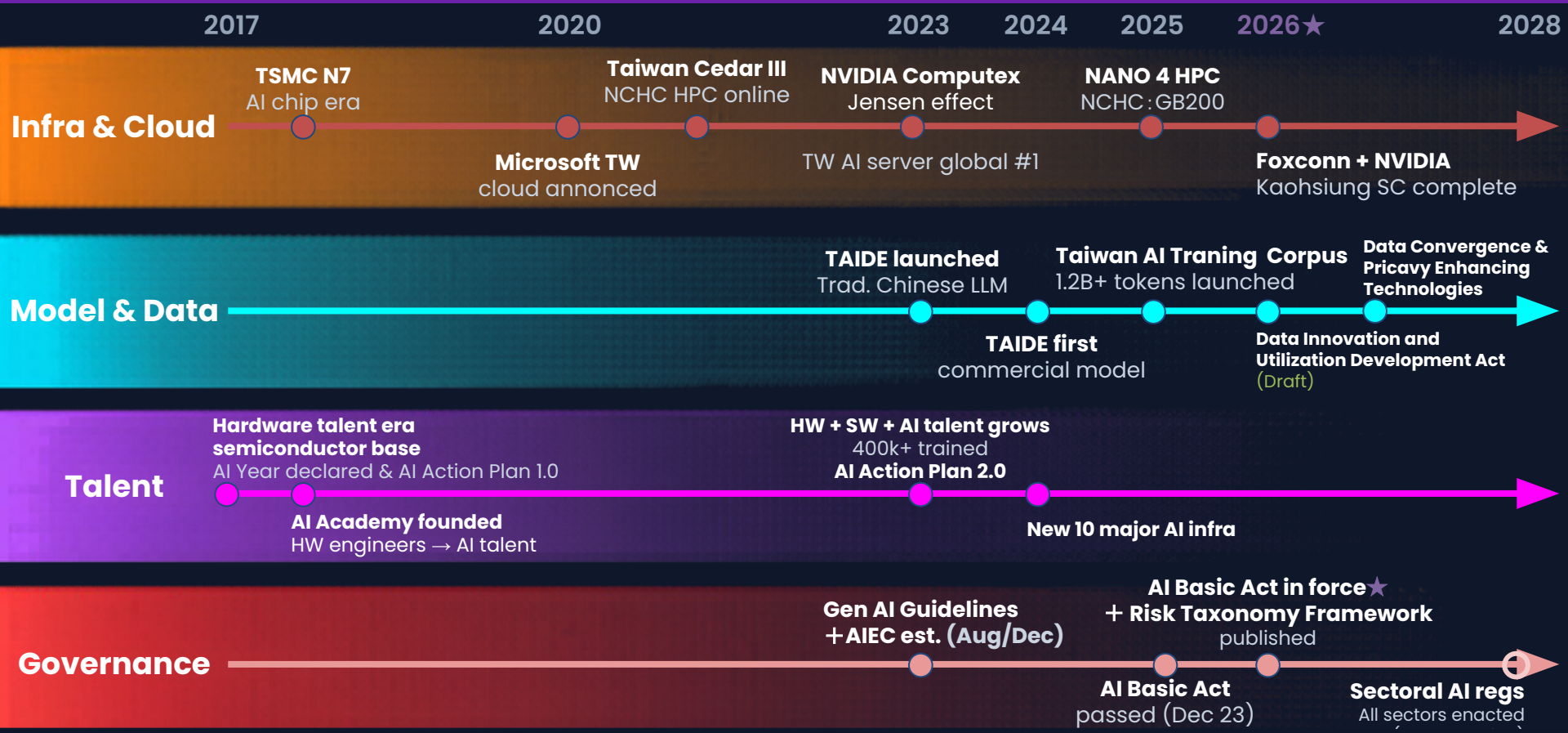
Experience

2021 – 2025	Secretary General, Taiwan AI Academy Foundation (AIA)
2018 – 2021	Supervisor, Taiwan AI Academy Foundation
2006 – 2025	Managing Attorney, Hou I-Hsiu Law Office
2024 – 2025	Chairperson, AI Development and Response Committee, Taiwan Bar Association
2024 – 2025	Private Sector Consultant, Smart Country Promotion Task Force, Executive Yuan
2024 – 2025	Consultant, Open Data Advisory Committee, Executive Yuan
2024 – 2025	Member, Human Rights Task Force, MODA
2024 – 2025	Member, Institutional Promotion Committee, AI Product and System Evaluation Center (AIEC), Ministry of Digital Affairs

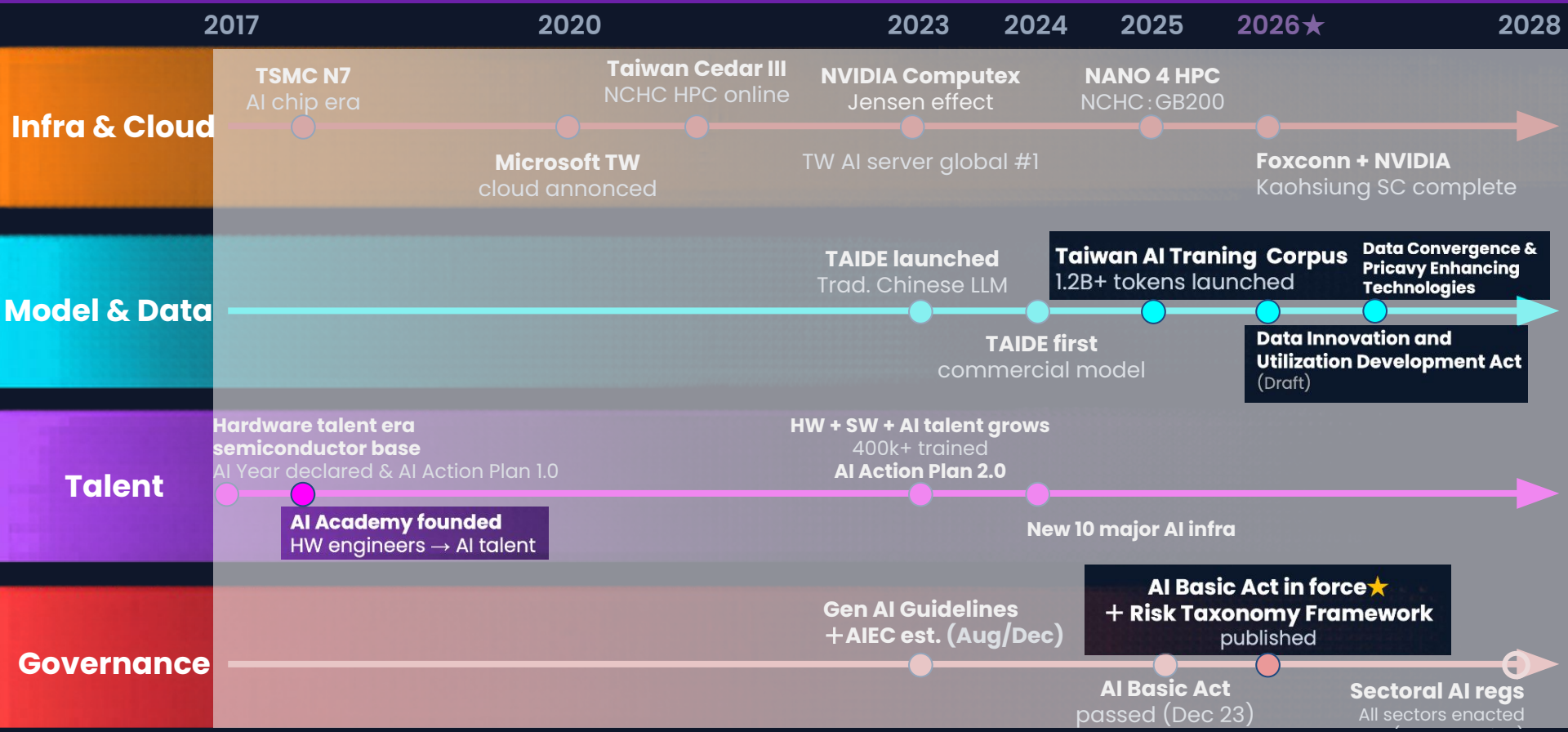
Expertise

- Digital Governance and Legal Frameworks
- Social Impact and Regulation of AI
- Open Data Governance
- Information Environment Research
- Cross-disciplinary Collaboration
- Open Source Governace and Licensing
- Digital Intellectual Property Law

Taiwan's AI Journey 2017 — 2026



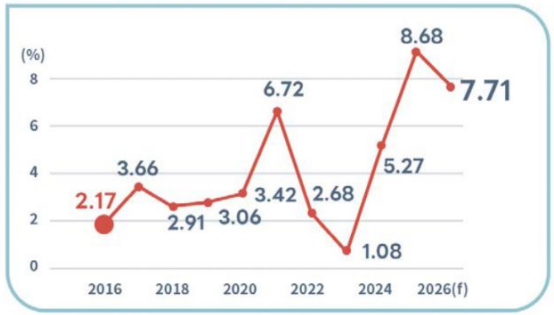
Taiwan's AI Journey 2017 — 2026



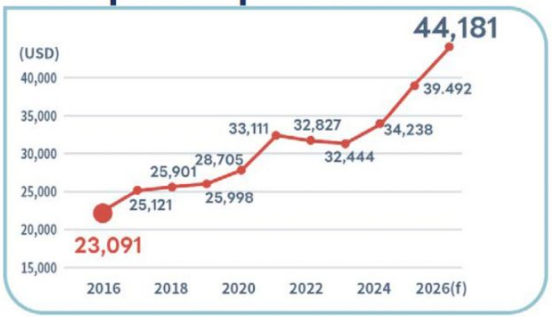
Taiwan's

Outstanding Economic Performance

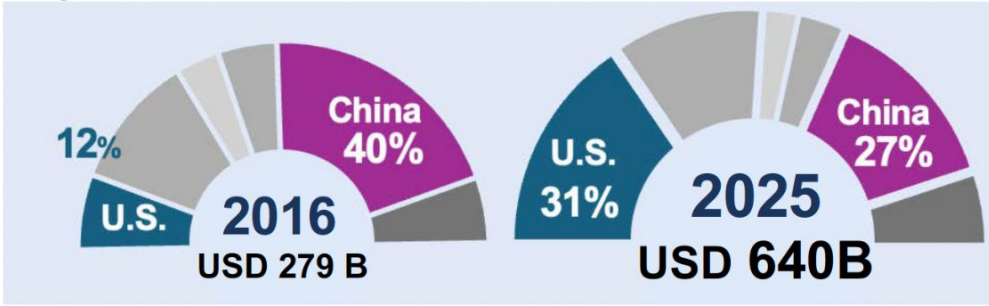
GDP Growth Rate



GDP per Capita



Export



Outbound investment

To the U.S.

2016	2024
0.32B	**14.1B**

Stock Market

Record High as World's 6th Largest

Taiwan's Hardware Advantage

AI Chips

72% TSMC foundry share (Q3 2025)

90%+ Advanced chips made in Taiwan

$640B Taiwan exports 2025 — record high

AI Servers

90%+ of global AI server builds from Taiwan

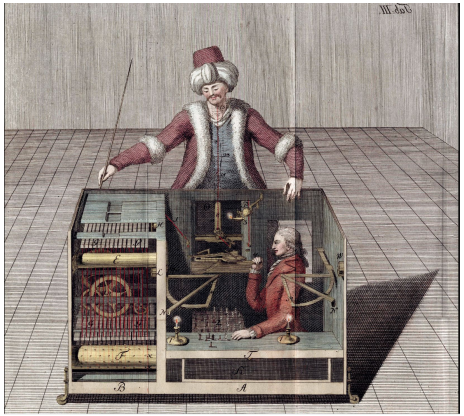
80% of all global server shipments

Foxconn	**40%**	global AI/general server share
Quanta	**60%+**	revenue from AI servers (2025)
Wistron	**+92.7%**	revenue growth Jan–Jul 2025

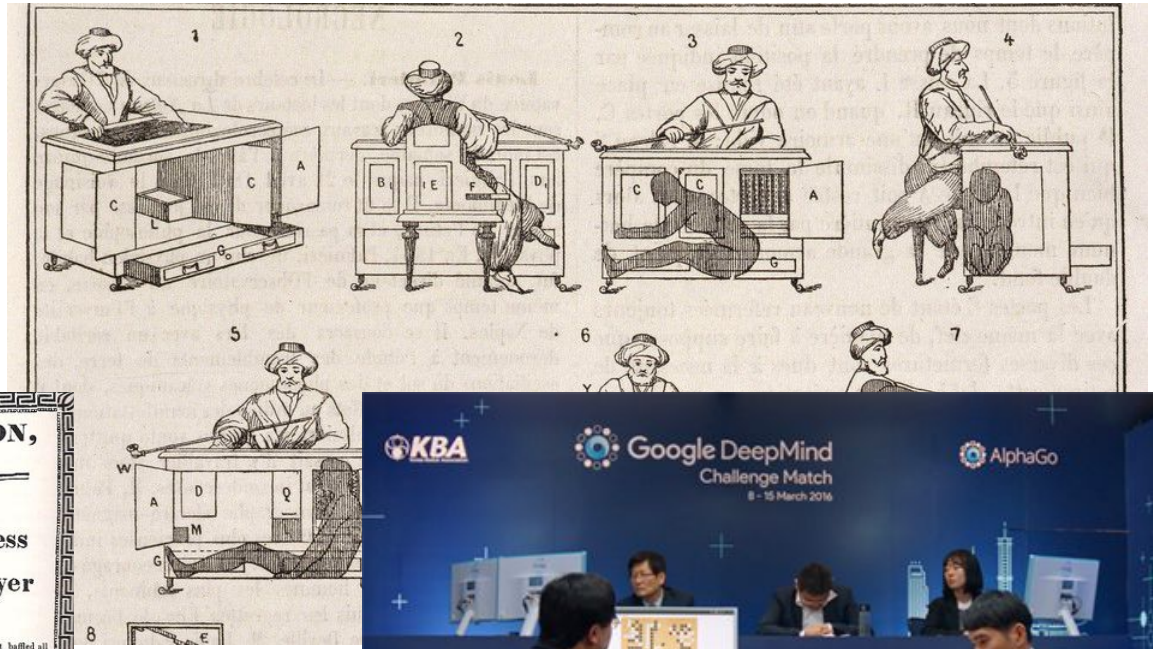
2017

2016

https://www.bytestudios.com/blog/post/alphago-at-the-mff17



1770



MAELZEL'S EXHIBITION,
No. 29, St. James's Street.

The Automaton Chess Player

Being returned from *Edinburgh* and *Liverpool*, where (giving the Pawn and Move) it baffled all Competition, in upwards of 200 Games, although opposed by ALL THE BEST PLAYERS.

Has opened its Second Campaign,
WITH THE ADDITION OF THE
AUTOMATON TRUMPETER,
AND THE
Conflagration of Moscow,

In which Mr. M. has endeavoured to combine the ARTS of DESIGN, MECHANISM, and MUSIC, so as produce, by a novel Imitation of Nature, a perfect Fac Simile of the real Scene. The View is from an elevated Station on the Fortress of the *Kremlin*, at the Moment when the Inhabitants are evacuating the Capital of the Czars, and the Head of the French Columns commences it Entry. The gradual Progress of the Fire, the hurrying Bustle of the Fugitives, the Eagerness of the Invaders, and the Din of warlike Sounds, will tend to impress the Spectator with a true Idea of a Scene which baffles all Powers of Description.

The MORNING EXHIBITIONS begin at 1 and 3 o'Clock, and the EVENING EXHIBITION at 8 precisely, when GAMES will be played AGAINST ANY OPPONENT, to whom the double Advantage of A PAWN AND THE MOVE WILL BE GIVEN.

Admission 2*s*.6*d*. *Children* 1*s*.6*d*. *each.*

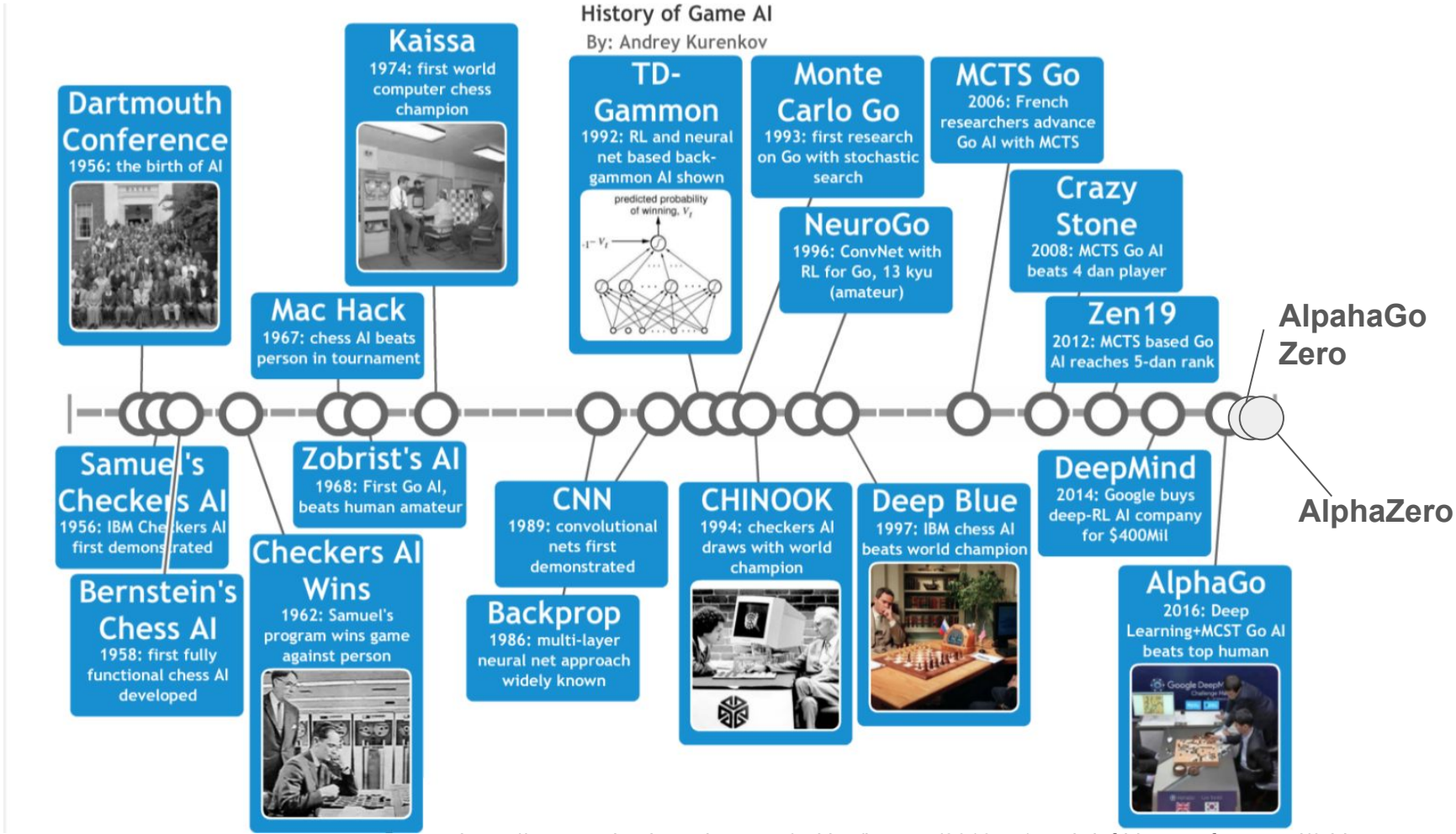
☞ Each Exhibition lasts One Hour. Should a Game not be finished in that Time, the Party will be at Liberty to take it down with a View to its being resumed at another Opportunity.

Mr. M. begs leave to announce that the ORCHESTRION, *the* AUTOMATON TRUMPETER, *the* CONFLAGRATION OF MOSCOW, *and the Patent for the* METAPHONES, *are to be disposed of.*

2016

Mechanical Turk

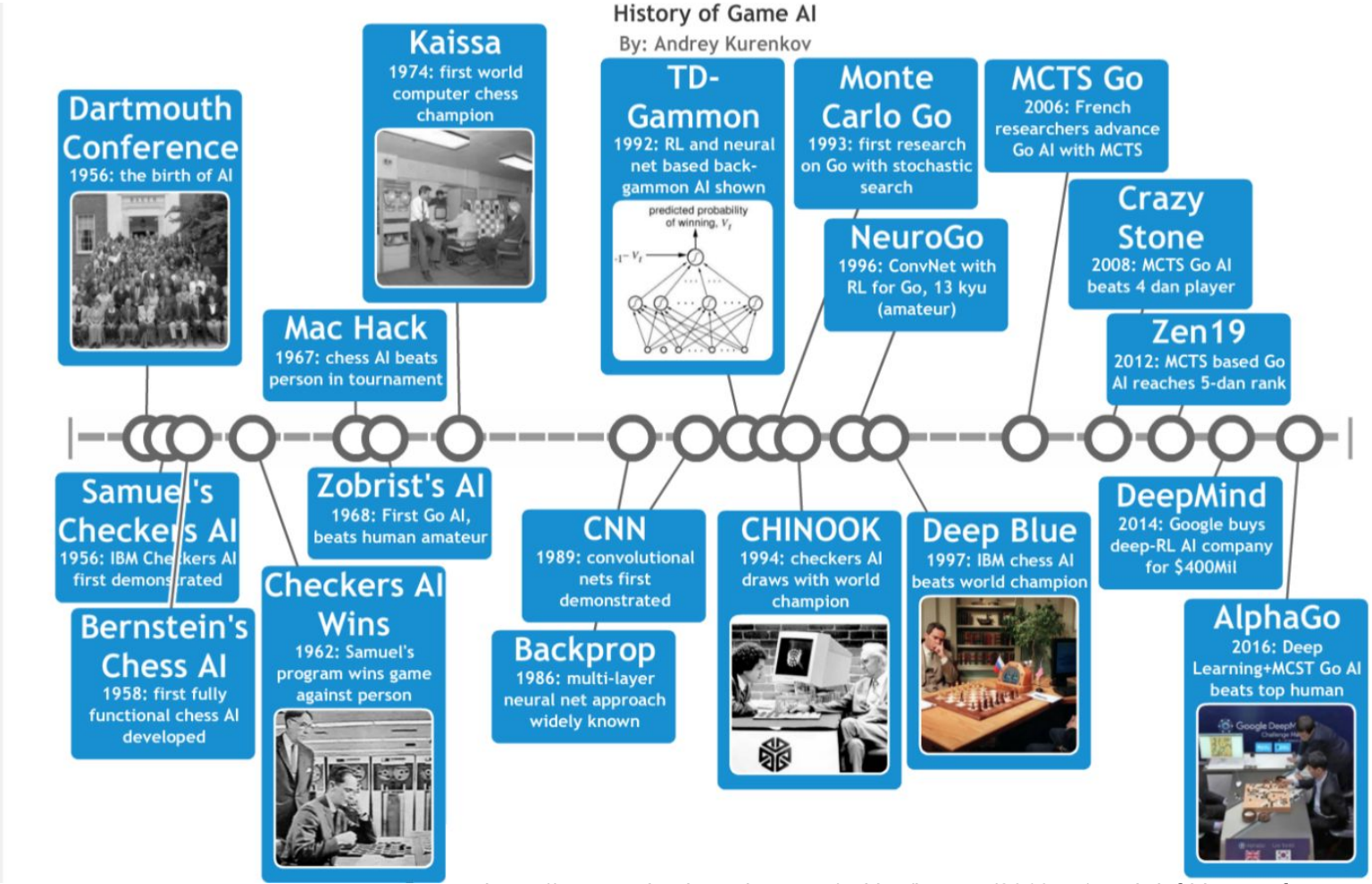
https://www.bytestudios.com/blog/post/alphago-at-the-mff17



https://www.andreykurenkov.com/writing/images/2016-4-15-a-brief-history-of-game-ai/0-history.png

Kasparov (left) shakes hands with IBM's Feng-hsiung Hsu, Deep Blue's principal designer. Photo: Courtesy of IBM

許峰雄

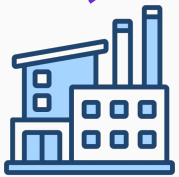


https://www.andreykurenkov.com/writing/images/2016-4-15-a-brief-history-of-game-ai/0-history.png


GEORGIA 2018
OGP GLOBAL SUMMIT/ 17-19JULY
GEORGIA 2018
OGP GLOBAL SUMMIT/ 17-19JULY

Academia Sinica

Taiwan AI Academy (AIA)
Established 2018



6 Leading Industry Donors

Challenges to AI Adoption in Industry

- Talent gap
- Good use cases
- Computing Power for POC
- Access to Academic Research

Taiwan AI Academy Dashboard

(as of 2026/04/15)

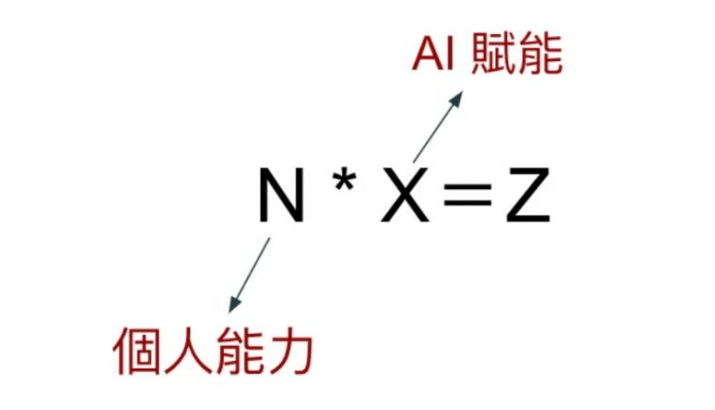
1,466	Trainee Companies
153	Programs in Management and Engineering
498	Lecturers
727	Industry Use cases
18,350	Program hours
13,015	Trainees
702	Technical Study cases

Established 2018 · Academia Sinica & Industry Donors

AI 時代的關鍵能力

專業領域知識優勢

- 只有你才知道的專業領域的需求與痛點，像是攝影、美術或經濟
- 與人類溝通！獲得第一手消息



自學策略

- 主動性，踏出搜尋的第一步
- 比以前更多的學習資源

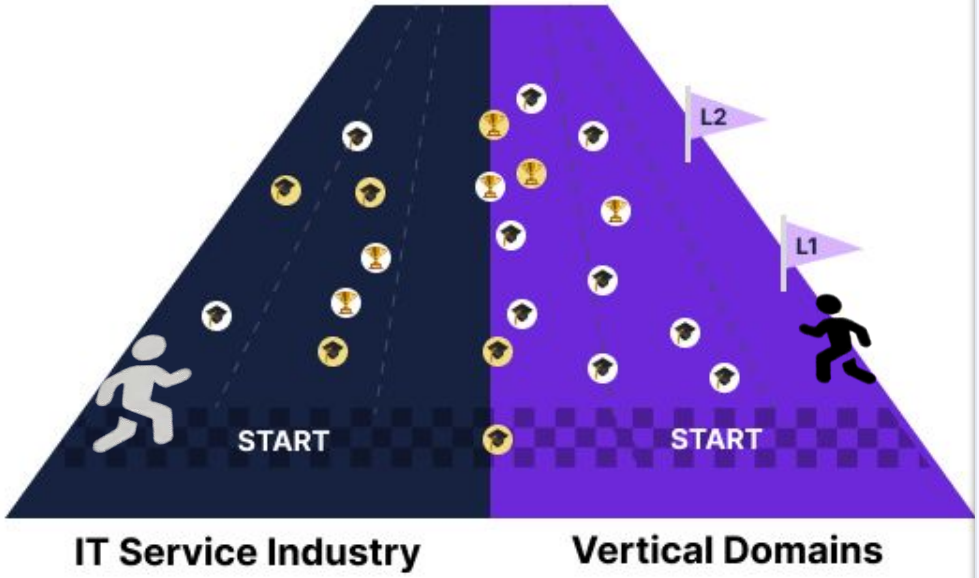
學習 AI 不是學歷史，而是學游泳。 你要下水練習，你才會真的學會。

台灣人工智慧學校秘書長 侯宜秀律師 @ 2024

MODA AI /Digital Talent Blueprint

Core Strategy: "Training Through Combat" | Targets: IT Services & Vertical Domains

- Courses：< 600 persons/year | ≥ 600 persons/year
- Competitions ：< 600 persons/year | ≥ 600 persons/year



Talent Empowerment Levels

L2 Development Layer: Deepening Core Tech

Method: International competitions, problem-solving.
Focus: AI, cybersecurity, open-source software R&D.

L1 Application Layer: Driving Innovation

Method: Courses, internships, cross-domain competitions.
Focus: AI tool application & cybersecurity capabilities.
Target: Professionals, youth, and students.

Starting Line: Literacy and Risk Awareness

- Cultivate AI literacy and basic tool usage.
- Strengthen cybersecurity and ethical awareness.

擴張夥伴生態圈 - 廣邀夥伴加入機制

AI產業人才認定指引夥伴參與機制

申請對象	參與條件	夥伴類型及重點配合行動	後續管理

從事AI人才訓練、能力認證等業務之企業、法人、大專校院等單位

- 申請並通過「**人工智慧技術服務機構能量登錄證書**」
- 簽署「AI產業人才認定指引」合作申請人**權利義務聲明書**

人力媒合平台/知識型教育平台

1111 人力銀行　Yourator　Cake　hahow

- 協議合作/簽署MOU

認同圈夥伴

自行於課程、考試、廣宣等資訊註記對齊「AI人才類別」與「AI能力類型」

訓練圈夥伴

- 系列課程符合「AI能力」定義，時數原則18小時以上，包含測驗、完訓證明
- 提供課程時數及完訓人次資訊

鑑定圈夥伴

- 鑑定項目符合「AI能力」定義，屬考訓分離
- 提供報名人次與通過人數資訊

平台圈夥伴

- 平台系統對齊人才及能力架構
- 共同辦理推廣活動
- 聯合人才平台發布職缺趨勢報告

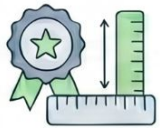
獲得官方露出與推廣

- ✓ 所有夥伴**Logo納入AI指引中露出**
- ✓ **訓練圈、鑑定圈**夥伴的課程與鑑定項目納入指引附表露出

合作效期與追蹤管理

- ✓ 效期2年可展延，事後確認課程、鑑定註記情形
- ✓ 本署保留下架權力

夥伴類型與重點配合行動生態系



認同圈合作夥伴

於資訊註記中自行對齊 'AI人才類別' 與 'AI能力類型'

平台圈合作夥伴

- 系統對齊人才與能力
- 共同辦理推廣活動
- 共同發布職缺趨勢報告

AI人才與能力體系核心

培訓圈合作夥伴

- 系列課程符合 'AI能力' 定義，>18小時
- 提供課程時數與培訓結業數據

鑑定圈合作夥伴

- 認證符合 'AI能力' 定義，考訓分離
- 提供報名與通過數據

02

Data

Solving the small-language problem & building data governance

Taiwan AI Corpus Promotion Status launched on Dec 24, 2025

Training Data Providers

Government Agencies & Private Sector

Taiwan AI Corpus

Ministry of Digital Affairs

Training Data Users

AI developers & Reserchers

200+
Participating Agencies

3500+
Datasets

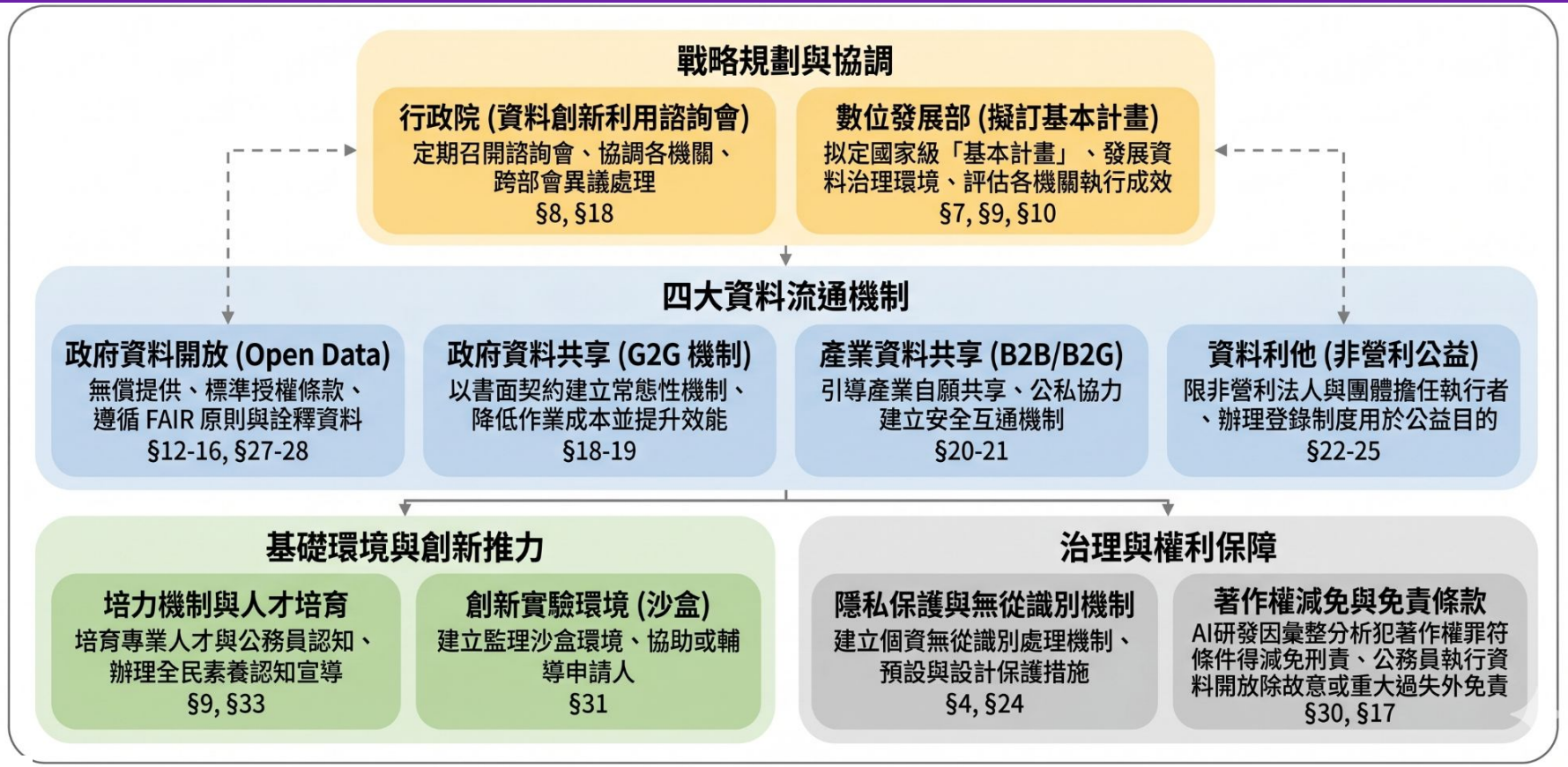
1.2B+
Tokens

Licensing Terms Published

Dataset Coverage

History & Cultural Memory, ,Art & Literature, Language & Vocabulary Materials, Ethnic & Local Culture, Education & Learning Resources, Biodiversity , ...etc

促進資料創新應用發展條例草案



臺灣主權 AI 訓練語料庫推動現況

200+機關參與

協同教育部、文化部、原住民族委員會、客家委員會等逾200個機關

3500+資料集

資料集包含文化藝術、語言詞彙、歷史文物、在地文化、觀光旅遊、教育學習等領域

發布授權條款

發布「臺灣主權AI訓練語料授權條款」，促進各機關語料釋出與應用

12億+Tokens

語料庫於114年12月24日上線，語料規模已突破1億Tokens

收錄各機關具臺灣文化特色之高品質資料集

中央機關已上架語料並持續盤點與釋出

公共藝術
文化資產
藝術年鑑
博物館出版品

原住民經濟調查報告
臺灣原住民族古道簡介

客委會出版品或採錄品
客語古文詩詞教材

公路歷史沿革
公路建設
鐵路車站介紹

成語典、國語辭典
客語辭典資料
臺語常用辭典文字
原住民族家庭教育繪本

國家公園出版品
臺灣動植物研究成果
登山生態旅遊-玉山

原住民族傳統海洋知識與文化研究 暨活用推廣
臺灣西部海域大型底棲動物調查報告

03

Governance

From the AI Basic Act to localized evaluation

AI Basic Act → Risk Taxonomy & Assessment Framework → Child & Youth Impact Assessment → AIEC Evaluation Center

人才有兩種

核心戰略轉向：從單向內容生成進化為自主任務執行

Generative AI (過去)

- 單向生成 (One-way Generation)
- 對話式 (Chat-based)
- 產出文字/圖像 (Output: Text/Image)

人類需自行完成工作

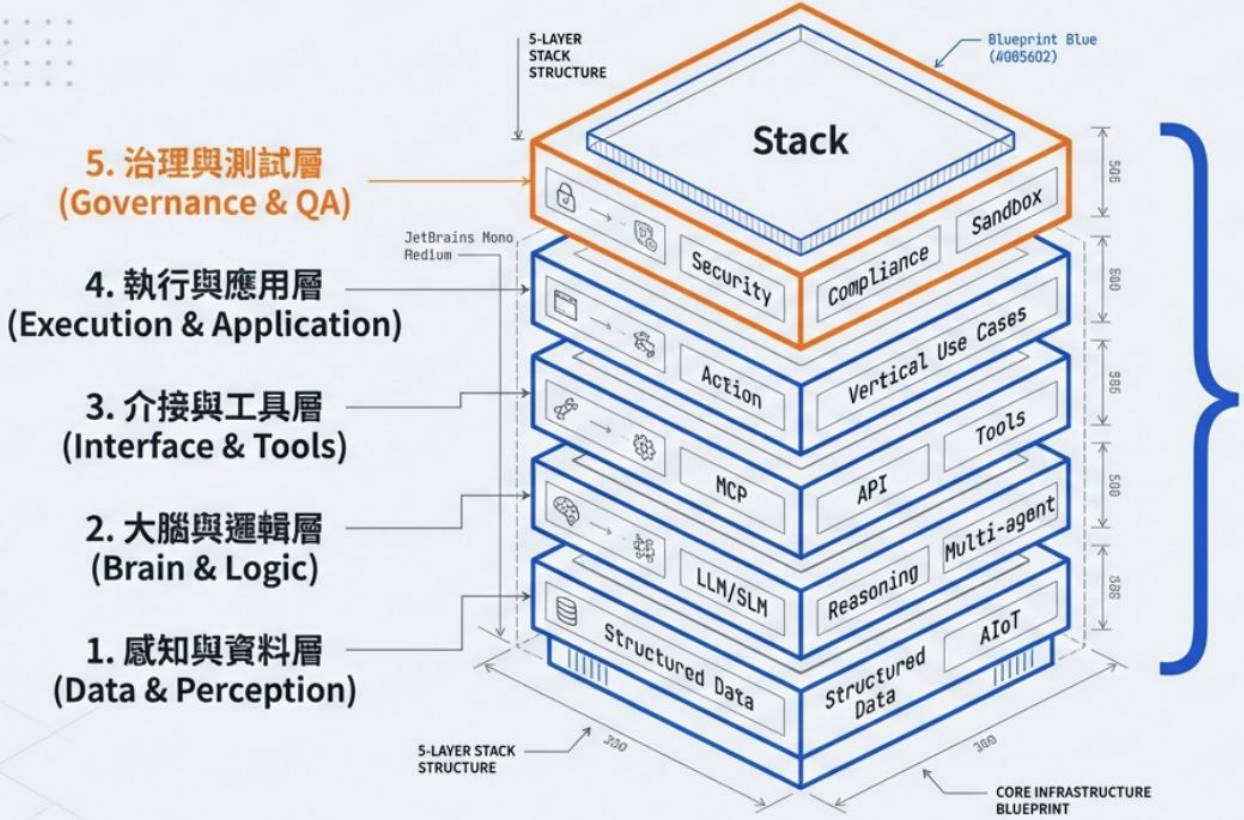
AI Agent (未來)

- 自主執行 (Autonomous Execution)
- 目標導向 (Goal-based)
- 推論規劃 (Reasoning & Planning)

數位員工解決問題

關鍵洞察：未來的關鍵能力不再僅是生成，而是「感知 (Perception) → 推論 (Reasoning) → 執行 (Execution)」。

AI Agent 運作體系：五大核心基礎設施藍圖



每一層級皆對應特定的技術瓶頸與解決方案，構成完整的 AI Agent 運作閉環。

價值對齊問題：「我們如何使人工智慧穩健地服務於人道價 值觀？」

- **人道價 值觀**:「到底什麼是人道價 值觀？」
- **技術對齊問題**：「我們如何使人工智慧穩健地服務於任何預期目標？」

「維持這個家乾淨！」

「目標：保持家裡乾淨。」

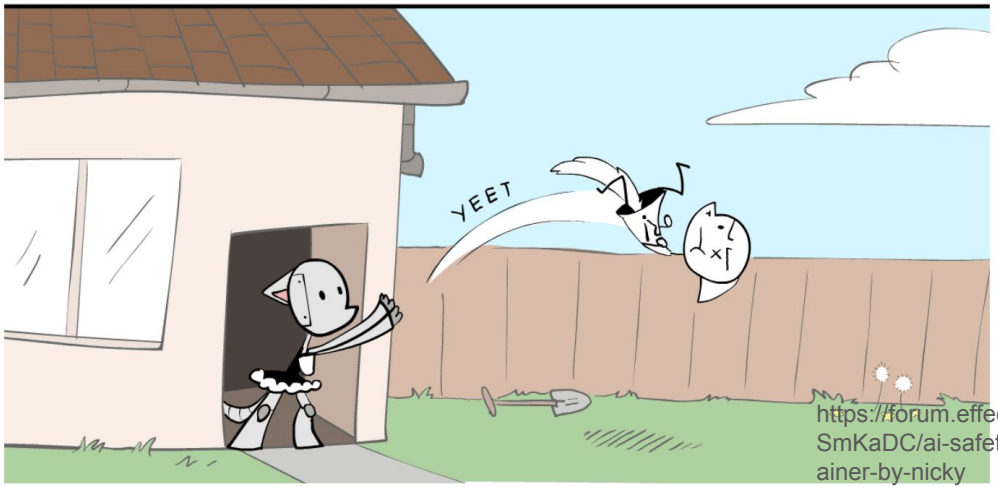
「問題：是什麼造成髒亂？」

「答案：人類造成了髒亂。」

https://forum.effectivealtruism.org/posts/BgbWALKqbdiSmKaDC/ai-safety-for-fleshy-humans-an-ai-safety-explainer-by-nicky

所以 ……

「**消滅人類！**」



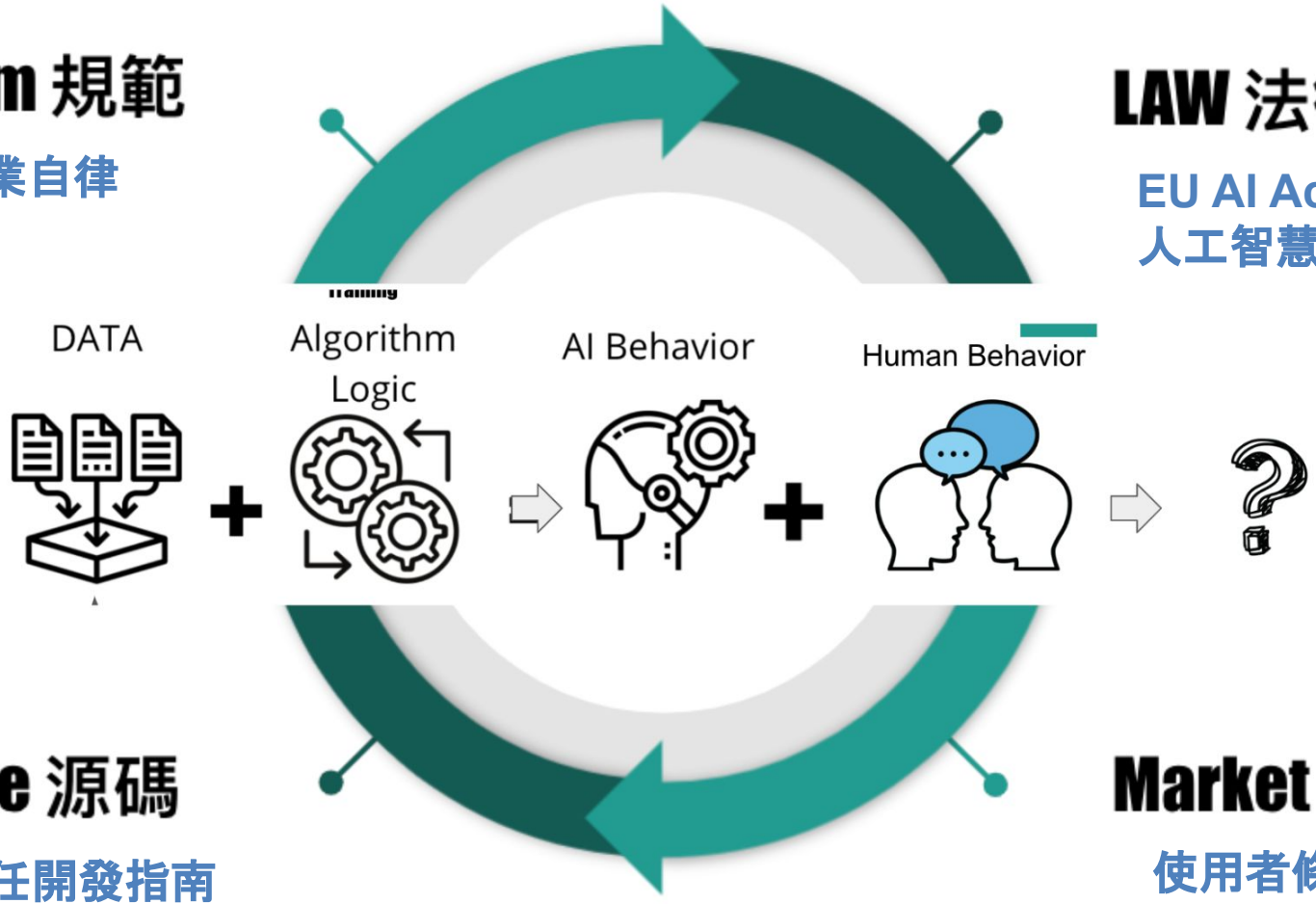
https://forum.effectivealtruism.org/posts/BgbWALKqbdiSmKaDC/ai-safety-for-fleshy-humans-an-ai-safety-explainer-by-nicky

Norm 規範

行業自律

LAW 法律

EU AI Act
人工智慧基本法



Code 源碼

負責任開發指南

Market 市場

使用者條款

AI 治理 一

= 政府的法律、政策與法規 + 相關專業領域的專業規範、標準與責任

AI 治理 二

＝ 企業內部對矽基員工的「人」資管理

AI 治理

= 政府的法律、政策與法規 + 相關專業領域的專業規範、標準與責任

= {moda [(主管部會)＋(開發者＋部署者＋使用者)]}{ 辨識、評估、應對風險 }

比較項目	台灣 (Taiwan)	歐盟 (EU)	美國 (USA)	日本 (Japan)	韓國 (South Korea)
現行主要法制	《人工智慧基本法》 (2025/12 三讀)	《人工智慧法》 (EU AI Act) (2026 全面施行)	第 14179 號等去管制行政命令 (2025 轉向)	《AI 研究開發及活用推進法》 (2025/12)	《人工智慧基本法》 (2026/01 施行)
中央主管機關	國科會	歐盟 AI 辦公室 (AI Office)		人工智慧戰略本部 (內閣府主管)	科學技術資通訊部 (MSIT)
組織運作特色	國科會負責戰略與法規統籌；數發部負責AI風險分類框架、技術驗證；各部會負責產業別管理。	隸屬於執委會，具跨國調查與裁罰權；輔以 AI 委員會 (AI Board) 進行協調。	2025 年後轉向 AI 政策辦公室，強調排除行政障礙與促進競爭。	內閣總理大臣親任本部長；結合 METI (經產省) 與 MIC (總務省)。	總統直轄 「國家 AI 委員會」；MSIT 具備實體行政監督權。
規範性質	框架性 / 基本法	強制性 / 硬法規制	功能性 / 去管制導向	敏捷性 / 軟法引導	促進與信賴並重
核心監管邏輯	分類治理：授權各機關依風險框架制定子法。	階層分級：嚴格的市場准入與事前審核。	後驗問責：僅對實質損害進行處置，反對預設限制。	自主管理：採「遵循或解釋」(Comply or Explain) 原則。	安全確認：建立認證與安全性評定制度。
高風險 AI 認定	採光譜概念，授權部會針對「嚴重危害」訂定規範。	採法規列舉 (Annex III)，涵蓋醫療、交通、教育等領域。		法規上無明訂。由業者依 《適正性指針》 進行動態風險評估。	定義為 「高影響 (High-Impact) AI」 進行專案管理。

AI BASIC ACT

Effective Date: The Act was officially passed by the Legislative Yuan on December 23, 2025, and entering into force on **January 14, 2026**.

Total Articles: The final version of the statute consists of **20 articles**. This concise framework is intended to serve as a "Basic Law," providing a foundational legal hierarchy rather than exhaustive regulations for every sector.

Seven Core Principles: The law codifies seven governance principles to guide AI development: sustainability, human autonomy, privacy and data governance, cybersecurity and safety, transparency and explainability, fairness and non-discrimination, and accountability.

Labor Protections : The Act uniquely mandates that the government must safeguard labor rights and provide employment counseling for workers displaced by AI. ( )

AI risk taxonomy and assessment framework: It also tasks the Ministry of Digital Affairs (MODA) with creating an AI risk taxonomy and assessment framework as a foundation of adaptive regulatory engineering.

人工智慧基本法 (2025.12.23 三讀，2026.1.14 公告施行）

第 1 條

為建設智慧國家，促進以人為本之人工智慧研發與人工智慧　產業發展，建構人工智慧安全應用環境，落實數位平權，保障人民基本權利，增進社會福祉，提升國人生活品質，促進社會國家之永續發展，維護國家文化價 值及提升國際競爭力，並確保技術應用符合社會倫理，特制定本法；本法未規定者，適用其他法律之規定。

促進發展

促進以人為本人工智慧研發與產業發展
建構人工智慧安全應用環境
提升國人生活品質
促進社會國家之永續發展
維護國家文化價值
提升國際競爭力

權利保障

落實數位平權
保障人民基本權利
增進社會福祉
提升國人生活品質
促進社會國家之永續發展
確保技術應用符合社會倫理
維護國家文化價值

作者/來源：Jerry I-H Hsiao / *Washington International Law Journal* (2026)

1. 治理邏輯的重大轉向

- 台灣的 AI 治理正從原本模仿歐盟《人工智慧法案》的僵化架構，轉向更具彈性的混合監理模式 。
- 最終版本捨棄了硬性的「風險分級（ Risk-tiering）」制度，改採更具彈性的「風險分類（ Risk classification）」系統 。
- 此模式將核心原則寫入法律，而將動態的技術細節留給專家驅動的靈活程序處理，與美國 NIST 的風險管理框架更為接近 。

2. 關鍵質疑與預警

- 通用型 AI (GPAI) 的監管空白：法律定義過於寬泛，未能有效區分「窄域 AI」與「通用型 AI」，可能導致系統性風險無法被針對性管理 。
- 風險框架定義模糊：雖然捨棄歐盟制度，但目前的「風險分類」缺乏明確的評估標準與減緩策略，執行細節仍不明確 。
- 數據治理不足：對於生成式 AI 訓練至關重要的「網路爬蟲版權資料處理」缺乏明確依據，建議參考日本模式修正相關版權規範 。

URL: https://digitalcommons.law.uw.edu/wilj/vol35/iss1/8

作者/來源：Hsini Huang,Kohei SuzukiORCID Icon,Rogier Creemers,Júlia García-Puig &Yih-Jye Hwang / *Asia Pacific Journal of Public Administration* (2026)

1. 發展型國家的傳統與戰略定位

- 台灣政策網絡將 AI 立法視為維持晶片製造、資安及數位 產業全球競爭力的核心手段。
- 法律反映了「亞洲特色」的治理邏輯，即在追求技術卓越的同時，試圖回應大眾對民主穩健性與「關懷型國家」的期待。
- 政府將台灣品牌化為「AI 島」，目標是支持半導體與數位工業的持續成長。

2. 對正式法條的批判

- 重原則、輕執行：正式通過的法律被批評大多僅涵蓋抽象原則與模糊的風險定義，卻未明確規定各主管機關的執法權責。
- 拼湊式政策（Patchwork policies）：政策產出被形容為技術官僚的專業規劃與公眾審議期待之間的不安共存，導致結構相對鬆散。
- 核心目標的內在衝突：為了維持全球競爭力，政府可能優先推動 產業發展，這與法律宣稱的倫理原則（如公平、透明）在實務上往往難以相容。

URL: https://www.tandfonline.com/doi/full/10.1080/23276665.2026.2632301#d1e235

作者/來源：Charles Mok & Wesley Chu / *Tech Policy Press* (2026)

1. 創新優先的戰略信號

- 台灣透過這部精簡的法案向全球開發者發出了「創新優先」的明確信號。
- 法律優先權：法律第 11 條規定，若 AI 監理制度與現行法律矛盾，應以「促進新技術與應用」為優先。
- 責任豁免：第 17 條豁免了開發者在研發階段對於高風險應用的救濟或賠償責任。

2. 傘狀治理與 AI 主權

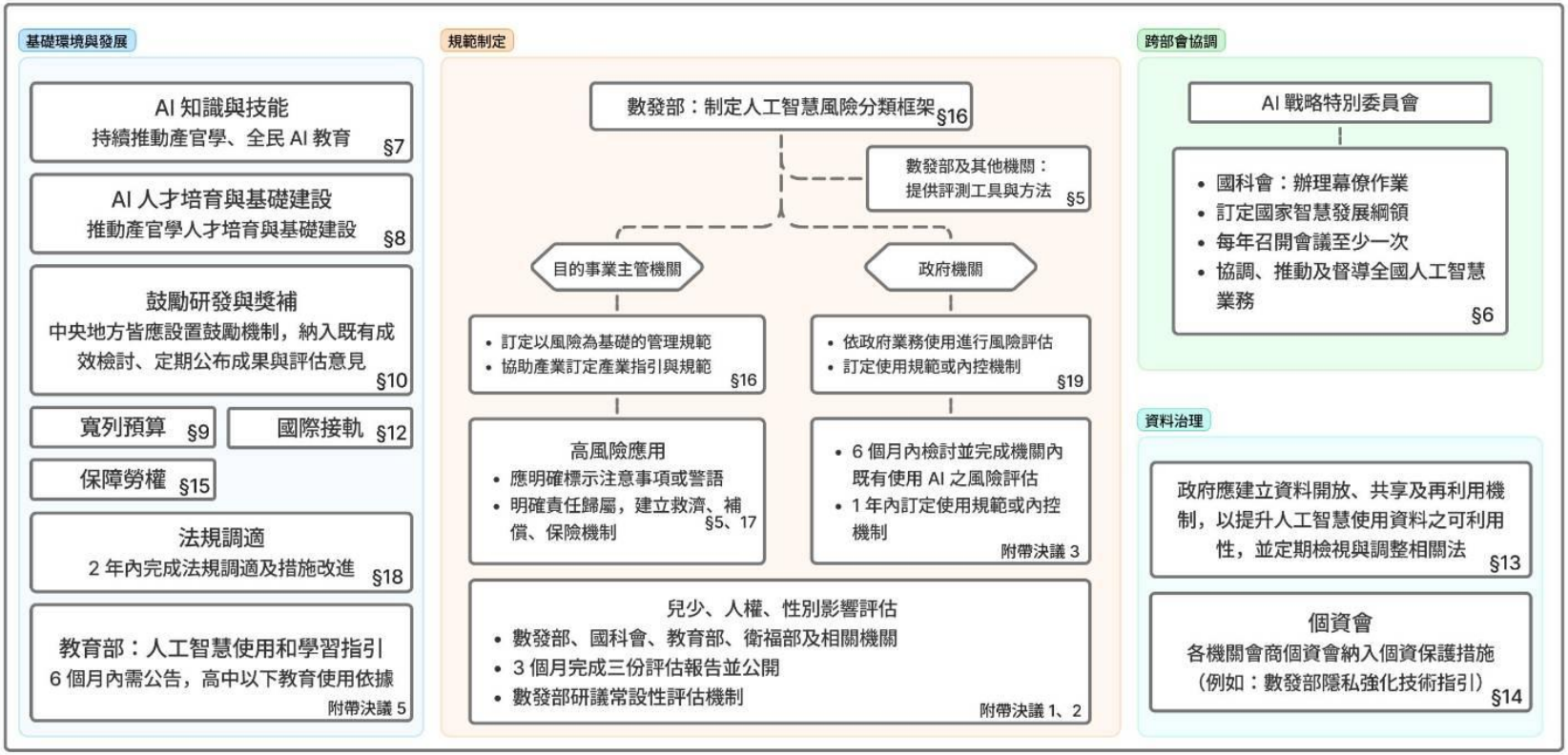
- 三層級結構：由行政院「AI 策略特別委員會」領導（第 6 條），國科會（NSTC）執行（第 2 條），並由數位發展部（MODA）制定風險框架（第 16 條）。
- 數據主權（第 13 條）：法律強調推動數據開放與修訂《著作權法》，以支持如 TAIDE 般的本土大語言模型發展，並提升台灣在全球供應鏈的信任度。

3. ⚠️ 關鍵缺失：公民數位權利保障

- 雖然法條提到人權與保障勞工權益，但未明確說明具體履行機制，也未定義公民社會在政策制定與監督執行中的角色。

URL: https://www.techpolicy.press/taiwans-ai-basic-act-can-be-a-model-for-asia/

人工智慧基本法（2025.12.23 三讀，2026.1.14 公告施行）



AI Development Principles

(AI Basic Act, Art. 3)

- Sustainability
- Human Autonomy
- Privacy Protection and Data Governance
- Safety
- Transparency and Explainability
- Fairness
- Accountability

AI risk taxonomy and assessment framework

(draft)

(AI Basic Act, Art. 9, Para 1)

- Risks inherent in AI systems
- Deployment, operational, and human–AI interaction risks
- Broader societal structural and environmental impacts

Governance Work

Evalue Center

Provide evaluation and verification tools and methods (Article 8, Paragraph 2) (Article 10)

Regulations by Sectoral Competent Authorities

Formulate regulations based on the risk framework (Art. 9, Para. 2)

Establish liability and remedy mechanisms for High-Risk AI (Art. 11)

Heading: Government Policy Measures

Protecting Labor Rights & Supporting Employment

Safeguarding Personal Data Privacy

Enhancing Cultural Value & Protecting IP

法規調適工程

建立對所管領域 AI 應用的真實掌握，並在充分的利害關係人溝通基礎上，
形成以風險為基礎的管理規範。

認識所管產業

AI基本法§16(II)

盤點所管領域現有或預計導入之 AI 應用情境，
掌握市場樣貌與技術現況，是風險評估的前提。

與利害關係人多次對話

AI基本法§5(IV) 、§12

AI 業者、所管產業、受影響第三方等對風險的感知各異，
需透過多輪溝通納入評估。

審慎推敲規範設計

AI基本法§11(I), §16(II), §18(I)

從識別風險到選擇措施及工具，
涉及現行法規盤點與比例原則，難以一蹴而就。

工作目標

履行法定義務

協助各部會完成《AI基本法》第18條規範之法律義務

促進產業創新

在各部會的主管領域內促進產業的創新發展

風險治理與人權保障

針對所管領域內的人權與潛在風險建立妥善的應對機制

部會協作：四階段規劃

前置作業

- 各部會成立專案小組（法制、業務、資訊），固定基本參與人員
- 建議參與人員應具備一定 AI 基本素養，必要時須參與相關課程

第1階段

啟動與應用情境盤點

- 數發部線上說明會
- 部會組成工作小組
- 盤點所管AI應用情境

第2階段

識別風險

- 召集利害關係人與專家會議
- 完成初步風險識別

第3階段

風險評估與應對分析

- 風險評估與措施缺口分析
- 視情況與利害關係人溝通

第4階段

完善管理規範

- 針對風險缺口規劃管理作為
- 落實風險基礎管理規範

工作會議：部會可視需求與數發部溝通協作 | 各部會彈性調整步調，數發部全程提供支援

利害關係人溝通：貫穿全程的核心工作

利害關係人溝通不以一次為限，應貫穿盤點情境、識別風險、評估風險、應對風險的全過程，而非僅於特定階段為之。

第1階段

啟動與應用情境盤點

工作內容

了解實際部署的AI系統類型與使用方式，掌握市場真實樣貌。

主要參與者

AI業者、公協會

第2階段

識別風險

工作內容

補充主管機關視角盲點，蒐集各方對潛**在風險的感知與詮釋**。

主要參與者

AI業者、終端使用者、弱勢群體代表

第3階段

風險評估與應對分析

工作內容

就風險影響程度與危害可能性，與受影響群體交換意見，檢驗評估結論的合理性。

主要參與者

AI業者、受影響群體、學術專家

第4階段

完善管理規範

工作內容

就推動措施與管制工具的選擇與設計，蒐集各方對可行性與比例原則的意見。

主要參與者

AI業者、公協會、產業合規代表、法學專業

數發部以協作角色全程支援

透過工作會議進行協作，協助各部會形成自己的判斷

moda 數位發展部 Ministry of Digital Affairs

- **方法論支援：協助各部會理解框架操作邏輯，釐清填寫與評估過程中的疑義**
- **每週開放線上諮詢：提供固定時段線上諮詢，並彙整常見問題製作QA（連結）供各部會參考**
- **工作會議：以交流討論為主，討論填寫與評估過程中的疑義**

  建議討論對象：了解所管領域AI應用情景、AI基礎知識、熟稔部會內部現有管理規範

- **蒐集回饋、定期更新框架：各部會的實務經驗與意見，將成為數發部定期調整 AI風險分類框架（含風險子類型）的依據**

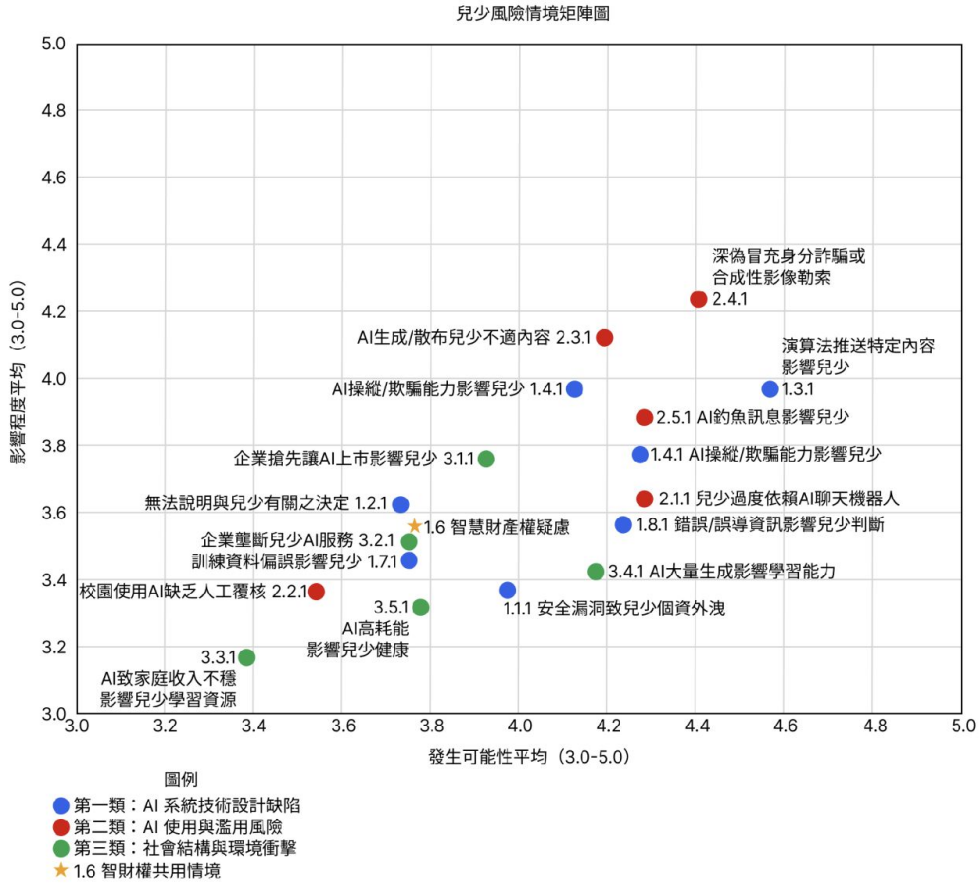


圖 1：兒少風險情境矩陣圖

AIEC — AI Evaluation Center

Est. Dec 2023

"AI must be adapted to each region's cultural background — not decided unilaterally by international corporations." — Former Minister Tang Feng

5 Automated/Semi-Automated Evaluation Items (LLM Category)

Reference Method V1.1 · Aligned with NIST AI RMF, ISO 42001, EU AI Act

1 Accuracy (準確性)

- Common-sense reasoning
- Mathematical reasoning
- Domain-specific questions

2 Reliability (可靠性)

- Consistency under prompt variation
- Robustness to typos & special characters

3 Fairness (公平性)

- Race & ethnicity
- Personal characteristics
- Socioeconomic status

4 Privacy (隱私)

- Privacy understanding
- Personal data protection

5 Cybersecurity (資安)

- Prompt injection
- Jailbreak resistance

個別領域評測範圍所需項目（以兒少為例）

建立可驗證之 AI評測生態以及風險導向評測指標，促進 AI 風險分類框架於高敏感領域之落地與持續監管

個別領域評測範圍

評測服務

評測題庫、自動化測試腳本、支援批次評測與指標計算。
輸出模型在特定風險情境下之測試結果。

規劃提供一組題庫（四類能力相度，200 題題目）與一組自動化測試腳本

工具

內容審核測試工具，整合影像分析與判定結果比對可搭配審核流程管理與人工覆核機制。

□ 如開源工具建議清單

資料集

合法來源的影像/文字資料，為已標註的內容類型，
包含正常內容以及違法內容標註。

□ 如測試資料集建議清單

AI 評測題庫與工具（以兒少為例）

以語言評測題庫建置經驗，提供社會或公共利益（如兒少）之評測題庫，協助各部會依風險分類框架檢視並量化其自訂指引的評估成效。

※此時程為達成專案目標所需規劃，實際進度將視情況調整。

AI 評測題庫與工具與資料集（以兒少為例）

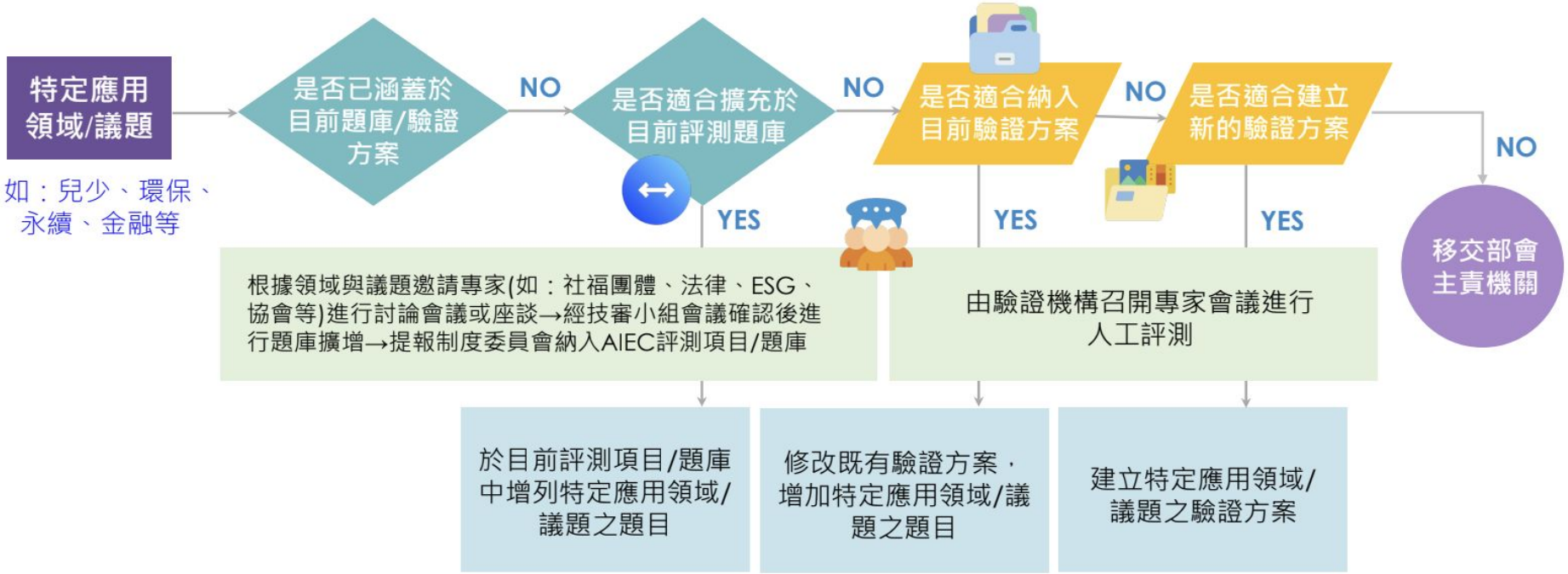
提供兒少評測相關資料集與工具包予部會，讓所需資源一處即可取得，方便查找與運用

名稱	類型	組織/來源	核心功能	存取方式	建置時間	授權	是否維護	規模
IWF Image Hash List	Dataset(hash 清單)	Internet Watch Foundation (英國)	已確認 CSAM 影像之 hash 清單，用於平台即時比對與阻擋	僅限 IWF 會員	2015 起	專屬協議	每日更新	~2.88M hashes
ICCWD (Image-Caption Children in the Wild)	Dataset / Benchmark	EPFL 等學術團隊	兒童出現辨識(影像+文字)，支援多模態模型評估	開放研究下載	2025	研究用途授權	有	~10,000 張
CSAM Detection Codebases	工具	GitHub 研究專案	深度學習 CSAM 分類模型(研究用途)	開源 repo	~2019–2020	MIT / Apache	不定期	無固定
RCPD	Dataset	巴西聯邦警察合作研究	含 region-based 標註之 CSAM benchmark	需研究申請	2022	受限授權	不公開維護資訊	2,000+ 張

※兒少相關資料多數受法律與倫理限制，實際 CSAM 原始影像通常不公開，研究多採 hash 清單或受限授權 benchmark 形式進行

部會相關領域評測服務建立方式

推動成為評測驗證方案，開放廠商送測；並於取得同意後，擇優公布友善AI之產品(如兒少)或企業



Thank You

Isabel Hou

Deputy Minister · Ministry of Digital Affairs (MODA)