立法院法制局專題研究報告　　編號：1774

本報告僅供委員參考

AI 人工智慧治理問題研析－以建構可信任 AI 為中心

法制局　李淑瓊　撰

中華民國 114 年 6 月

AI 人工智慧治理問題研析－以建構可信任 AI 為中心

目　次

摘　要

隨著人工智慧技術於全球快速擴展，其同步引發諸如資料隱私、演算法歧視、自動決策問責與倫理問題等風險，亦同時被重視。如何於促進創新與保障公共利益之間取得平衡，已成為各國AI治理機制之政策核心。本報告以「建構可信任AI」為主軸，分析人工智慧治理之國際趨勢、外國立法經驗及我國現行制度問題，提出未來立法應以發展導向為基礎，兼顧風險控管與產業支持、建置可信任AI之驗證環境、建立AI監理沙盒制度，同時發展主權AI，以維護我國語言文化、自主技術與國家安全等研析意見，並提出相關立法建議，俾供本院委員問政參考。

AI人工智慧治理問題研析－以建構可信任AI為中心

壹、 前言

隨著人工智慧（Artificial Intelligence，AI）技術的發展及其應用的普及，其所涉及之倫理、隱私洩漏、演算法的偏見、系統失能、濫權監控、深偽技術誤導，甚至侵犯基本人權等安全性與可靠性問題，逐漸成為人工智慧治理之重要議題。是以，建構可信任的AI（Trustworthy AI）逐漸成為國際社會於發展AI時的共通議題。世界主要國家（如：美國及歐盟）皆積極推動AI治理原則與立法機制，企圖在促進創新與保障公民權益之間取得平衡。相對而言，我國目前尚未制定AI專法或風險評估架構，使得民間業者在面對AI倫理與法律風險時，缺乏明確指引或合法標準。爰本報告擬檢視近年世界主要國家及國際組織所制定發布有關可信任AI之相關規範，並評估各國監管機制中值得借鏡之處，提出具體立法與政策建議，期能對我國建立具有前瞻性與可操作性的AI監管體系有所助益。

貳、 人工智慧的發展背景與可信任 AI 的概念

一、人工智慧的發展背景與造成的影響

人工智慧泛指由電腦模擬人類智慧的技術，涵蓋機器學習、深度學習、自然語言處理、電腦視覺等分支。自2010年代深度學習技術興起後，AI技術進入實用化階段，並於各領域展現驚人成效。國家科學及技術委員會預告制定之「人工智慧基本法」草案[1]第2條將「人工智

[1] 國家科學及技術委員會：預告制定「人工智慧基本法」草案，公共政策網路參與平臺，2024 年 7 月 15 日，網址：https://join.gov.tw/policies/detail/4c714d85-ab9f-4b17-8335-f13b31148dc4，最後瀏覽日期：2025 年 3 月 14 日。又行政院 2025 年 2 月已指示 AI 基本法草案改由數位發展部主政推動法案立法及續行條文解釋，參見：蘇思云，AI 基本法草案 數發部：在創新與風險中找平衡，中央通訊社，2025 年 4 月 23 日，網址：https://www.cna.com.tw/news/afe/202504230387.aspx，

慧」定義為「係指以機器為基礎之系統，該系統具自主運行能力，透過輸入或感測，經由機器學習與演算法，可為明確或隱含之目標實現預測、內容、建議或決策等影響實體或虛擬環境之產出。」[2]。AI系統雖為人類帶來效率與便利等正面效益，但亦產生決策黑箱、演算法偏好導致的歧視風險、個資與隱私外洩、決策透明度不足、監控濫用等負面影響。此外，AI系統的不透明性與不易問責性，時常引發倫理與法律爭議，這些挑戰迫使各界積極探討如何建立負責任且可信任的AI應用框架。

二、可信任 AI 的概念

歐盟人工智慧高階專家小組（High-Level Expert Group on AI, HLEG）於2019年4月8日所發表的「可信任AI倫理準則」（Ethics Guidelines for Trustworthy AI）中有助於理解可信任AI的概念，該準則提出可信任AI應具備3個要件及7個關鍵要求，準則說明如下[3]：

（一）可信任AI要件：

1.合法性（Lawfulness）：遵守所有適用法律與規範。

2.倫理性（Ethicalness）：符合倫理原則與價值。

3.穩健性（Robustness）：既從技術角度考慮，也考慮其社會環境。

（二）可信任AI關鍵要求：

1.人類的主導和監督：人工智慧系統應該賦予人類權力，使他們能夠做出明智的決定並維護他們的基本權利。同

最後瀏覽日期：2025 年 5 月 1 日。

[2] 國家科學及技術委員會，國家科學及技術委員會公告：預告制定「人工智慧基本法」草案，公共政策網路參與平臺，2024 年 7 月 15 日，網址：https://join.gov.tw/policies/detail/4c714d85-ab9f-4b17-8335-f13b31148dc4，最後瀏覽日期：2025 年 3 月 14 日。

[3] European Commission，Ethics guidelines for trustworthy AI，2019 年 4 月 8 日，網址：https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai，最後瀏覽日期：2025 年 3 月 28 日。

時，需要確保適當的監督機制，這可以透過人為干預（human-in-the-loop）、人為監控（human-on-the-loop）及人為控制（human-in-command）的方法來實現。

2. 技術穩健性和安全性：人工智慧系統需要具有彈性和安全性。它們必須是安全的，確保在出現問題時有備案計畫，並且具備準確、可靠和可重複性。這是唯一能確保即使是非故意之傷害也能被最小化甚至避免的方法。

3. 隱私和資料治理：除了確保充分尊重隱私和資料保護之外，還必須確保充分的資料治理機制，考慮到資料的品質和完整性，並確保對資料的合法存取。

4. 透明度：數據、系統和AI商業模型應該透明。可追溯機制可以幫助實現這一點。此外，應以符合相關利害關係人需求的方式解釋人工智慧系統及其決策。人類需要意識到他們正在與人工智慧系統互動，並且必須了解該系統的功能和限制。

5. 多元化、非歧視和公平：必須避免不公平的偏見，因為它可能產生多重負面影響，從弱勢群體的邊緣化到偏見和歧視的加劇。為了促進多元性，人工智慧系統應該讓所有人都能使用，無論其是否有任何障礙，並應在整個生命週期中納入相關利害關係人的參與。

6. 社會與環境福祉：人工智慧系統應該造福全人類，包括子孫後代。因此必須確保它們是具永續性並且對環境友善。此外，它們還應該考慮環境，包括其他生物，

並審慎評估它們對整體社會的影響。

7.問責制：應建立機制，確保人工智慧系統及其結果的責任和問責。可審計性使得對演算法、數據和設計過程的評估成為可能，在其中發揮關鍵作用，尤其是在關鍵應用中。此外，還應確保提供充分且易於獲得的補救措施。

參、 外國立法例及國際規範比較分析

一、主要國家 AI 治理規範

（一）美國

美國對人工智慧的監管是採取較為寬鬆的監管方式，尚未制定聯邦層級的 AI 專法，目的在鼓勵創新。雖然美國還沒有一個全國統一的法律框架來全面規範 AI，但仍有一些機構和政府部門提出了指導原則和框架來確保 AI 技術的可靠性、安全性和道德性，分述如下：

1.人工智慧應用監管指引[4]

美國白宮於 2020 年 11 月 17 日發布「人工智慧應用監管指引」（Guidance for Regulation of Artificial Intelligence Application），為聯邦機構就人工智慧的監管提供指引，以保持美國在人工智慧領域的領導地位。

2.人工智慧風險管理框架 1.0

美國國家標準與技術研究院（National Institute of Standards and Technology, NIST）於 2023 年 1 月 26 日公布「人工智慧風險管理框架 1.0」(Artificial Intelligence

[4] 該指引全文參見：Whitehouse，Guidance for Regulation of Artificial Intelligence Applications，2020 年 11 月 17 日，網址：https://www.whitehouse.gov/wp-content/uploads/2020/11/M-21-06.pdf，最後瀏覽日期：2025 年 3 月 13 日。

Risk Management Framework, AI RMF 1.0)，該自願性框架提供相關資源，以協助組織與個人管理人工智慧風險，並促進可信賴的人工智慧（Trustworthy AI）之設計、開發與使用[5]。該框架鼓勵組織建立 AI 風險治理策略，包括：明確組織責任與 AI 治理結構（Governance）、系統性風險評估與記錄管理（Map）、實施技術與非技術風險控制措施（Manage）及監測與改善機制（Measure）[6]。

3. 人工智慧監管指導原則[7]

美國聯邦貿易委員會（FTC）就保護消費者權利及如何發現與避免詐騙，發布許多建議及指導原則，2023 年發布「2023 年隱私與資料安全更新報告」明確表示人工智慧的應用須遵循既有的消費者保護法規，強調「AI 並不享有法律豁免」，FTC 也持續透過政策指引等指導原則，提醒企業開發與部署 AI 應該遵循「以人為本、公平、透明與負責任」的原則，確保不會對消費者的隱私與安全造成損害。

4. 人工智慧行政命令

美國前總統於 2023 年 10 月 30 日簽署「安全、可靠、可信賴的人工智慧開發與使用」（Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence）行政命令（下

[5] 陳箴，美國國家標準與技術研究院公布人工智慧風險管理框架（AI RMF 1.0），財團法人資訊工業策進會科技法律研究所，2023 年 4 月，網址：https://stli.iii.org.tw//article-detail.aspx?tp=1&d=8974&no=64，最後瀏覽日期：2025 年 3 月 28 日。

[6] NIST，AI Risk Management Framework，網址：https://www.nist.gov/itl/ai-risk-management-framework，最後瀏覽日期：2025 年 3 月 21 日。

[7] Federal Trade Commission，The Federal Trade Commission 2023 Privacy and Data Security Update，2024 年 3 月 21 日，網址：https://www.ftc.gov/system/files/ftc_gov/pdf/2024.03.21-PrivacyandDataSecurityUpdate-508.pdf，最後瀏覽日期：2025 年 5 月 1 日。

稱 AI 行政命令），此為美國第一個直接涉及人工智慧管理的行政命令，該行政命令要求美國公私部門以安全、可信賴的方式發展及使用 AI，並期望藉由全面性的行政命令，引導各領域思考如何應對 AI 風險[8]。惟此行政命令被認為阻礙 AI 創新，遭現任總統於 2025 年 1 月 23 日簽署「消除美國人工智慧領導障礙」（Removing Barriers to American Leadership in Artificial Intelligence）行政命令予以撤銷[9]，可以預期美國對人工智慧的監管策略將會重新調整。

有學者認為，美國的自由市場傳統強調個人主義和最小政府干預，這些都促使美國在 AI 監管上更側重於市場的自我調節和技術創新，主要以倡議性文件和分散監管的方式來面對[10]。此種監管方式可能有利於創新和技術突破，但亦有可能會犧牲部分隱私和安全性。

（二）歐盟

歐盟對人工智慧的監管較為積極，注重於保障民眾權益、維護社

[8] 臺灣 AI 卓越中心，美國白宮發布 AI 行政命令，網址：https://www.twaicoe.org/%E7%BE%8E%E5%9C%8B%E7%99%BD%E5%AE%AE%E7%99%BC%E5%B8%83AI%E8%A1%8C%E6%94%BF%E5%91%BD%E4%BB%A4，最後瀏覽日期：2025 年 3 月 5 日。AI 行政命令全文參見：總統辦公室於 2023 年 11 月 1 日 發布的總統文件，網址：https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence，最後瀏覽日期：2025 年 3 月 5 日。

[9] 川普總統於 2025 年 1 月 23 日簽署了第 14179 號行政命令「消除美國人工智慧領導障礙」（Removing Barriers to American Leadership in Artificial Intelligence），撤銷了拜登總統於 2023 年 10 月 30 日簽署的第 14110 號 AI 行政命令，該行政命令全文參見：Whitehouse，Removing Barriers to American Leadership in Artificial Intelligence，2025 年 1 月 23 日，網址：https://www.whitehouse.gov/presidential-actions/2025/01/removing-barriers-to-american-leadership-in-artificial-intelligence/，最後瀏覽日期：2025 年 3 月 5 日。

[10] 陳岩，人工智能大爆發後的監管命題：風險分類、路線分化與地緣政治陰影，BBCNEWS 中文，2025 年 3 月 14 日，網址：https://www.bbc.com/zhongwen/articles/ckgnvl7pwr3o/trad?at_link_id=252AB14A-02CA-11F0-BAA9-BE9226ED606D&at_campaign=Social_Flow&at_link_origin=bbc_news_trad&at_medium=social&at_campaign_type=owned&at_ptr_name=facebook_page&at_format=link&at_bbc_team=editorial&at_link_type=web_link&fbclid=IwY2xjawJEheVleHRuA2FlbQIxMQABHS6upP0vvfzn0yIkLSZswxUkz6cSZcfxQSsAORUwb3Gp93MNirmw0186lQ_aem_mNscyVHf5qjjbf3X5zfoag，最後瀏覽日期：2025 年 3 月 17 日。

會公平，並防止 AI 技術可能帶來的負面影響。主要監管法規有被譽為全球最嚴格的隱私和安全法的歐盟「一般資料保護規則」(General Data Protection Regulation，GDPR）及「人工智慧法」(Artificial Intelligence Act，AIA)，前者對 AI 的規範包括處理個人資料需獲個人明確同意，個人有權反對其個人資料被使用，也有權要求刪除[11]；後者定義了 4 種 AI 風險類別，並針對不同的類別，賦予不同程度義務及行為準則，且該法具有域外效力，亦即任何被歐盟國家公民使用的 AI 均需遵守[12]。

歐盟執委會(European Commission)在 2021 年 4 月提出「人工智慧法」草案。歐洲議會（European Parliament）於 2024 年 3 月 13 日正式表決通過人工智慧法。茲將該法中之風險分級監管措施及監理沙盒制度介紹如下：

1.風險分級監管措施

歐盟的人工智慧法規範了人工智慧的監管框架，將人工智慧系統的風險分為不可接受風險（Unacceptable Risk)、高度風險（High Risk)、有限風險（Limited Risk）以及最低風險或無風險（Minimal or no Risk）四個層級，並就不同風險層級的人工智慧系統，規定不同的監管方式，茲將不同風險層級的人工智慧系統之適用範圍及所受限制，概述如表 1：

[11] 歐盟「一般資料保護規則」全文參見：歐盟資料網頁，網址：https://gdpr-info.eu/，最後瀏覽日期：2025 年 3 月 5 日。

[12] 歐盟「人工智慧法」全文參見：歐盟官方公報，網址：https://artificialintelligenceact.eu/the-act/，最後瀏覽日期：2025 年 3 月 5 日。

表 1：歐盟人工智慧法之風險分級及監管措施表

風險等級	適用範圍	所受限制
不可接受風險(Unacceptable Risk)	嚴重影響基本權利與歐盟價值觀的AI應用。 例如：在車站、街道、商場等公共場所，使用即時遠距生物特徵（如臉部、指紋、虹膜等）識別。	嚴格被禁止使用
高度風險（High Risk）	對健康、安全及基本權利構成高度風險的AI系統。 例如：評估自然人社會福利與服務資格、評估消防與醫療緊急調度之優先度、測謊或具有類似功能之工具、能協助司法機關研究與解釋法律以及事實(fact)之人工智慧系統、能影響投票結果或投票權之行使的人工智慧系統及於社群媒體平台使用之個人化推薦系統等。	應建立風險管理機制，且該機制需在人工智慧系統的生命週期中持續運作與更新;須能自動記錄活動，確保可追溯性；具備適當的人類監督措施，可由人類進行監督，以減少風險;具備一定程度的準確性、穩健性、安全性和網路安全性，並在整個生命週期中有一致性的表現。進入市場前，亦必須向監管機構提供技術文件，包括預期目的、產品開發與風險管理等流程的詳細描述。
有限風險(Limited Risk)	具特定透明度的AI系統（已知的演算法、資料處理流程），此類AI系統存在一定的風險，但風險程度較低。 例如：聊天機器人，以及能創造出文字、圖像或影音等生成式AI系統(如ChatGPT)。	需滿足透明度要求。必須讓使用者及時且明確地知道自己正在跟AI互動，或正在觀看與使用AI創造的內容，並且揭露製作者、監督者等資

		訊。不能阻止執法機構檢測深度造假（deepfake）或進行刑事犯罪調查。
最低風險或無風險（Minimal or no Risk）	風險極低或不構成風險的 AI 應用。 例如：人工智慧電玩遊戲、垃圾郵件過濾器等。	不受監管，可自由使用。

資料來源：作者彙整自歐盟人工智慧法條文

歐盟人工智慧法透過風險分級制度，建立一致的監管框架，有助於促進成員國間的標準化與市場統一。該制度依據 AI 系統的風險高低，制定對應的監管措施，具備針對性與彈性。此外，風險導向能集中行政資源於高社會影響領域，提高監管效率。

然而，有學者指出風險分級制度的侷限性，例如：風險分級制度通常依賴可量化的指標來評估風險，惟許多與 AI 相關的危害卻難以量化（如歧視、隱私侵犯、尊嚴損害等）；又例如：風險分級制度可能將實際上的政策決策偽裝為技術決策，使得關鍵的價值判斷和政策選擇缺乏透明度和民主參與；另外，風險分級制度通常缺乏有效的反饋機制，無法從實際案例中學習並調整規範，導致制度僵化，難以適應快速變化的技術環境[13]。又在 AI 系統跨領域間應用時，單一風險分級可能無法充分反映其跨領域的綜合風險，導致執行困難。綜上可知，歐盟人工智慧法的風險分級監管方式有其優勢，

[13] Kaminski, Margot E.，〈The Developing Law of AI: A Turn to Risk Regulation. The Digital Social Contract: A Lawfare Paper Series〉，《U of Colorado Law Legal Studies Research Paper》，No. 24-5，2023 年 4 月 2 日，網址：https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4692562#，最後瀏覽日期：2025 年 5 月 1 日。

但也存在一些挑戰，尚需要透過具體實踐以調整及完善。

2. 監理沙盒制度

歐盟人工智慧法是全球第一部試圖全面規範人工智慧技術之立法，該法案與其他國家相較係採較嚴格之監管，惟條文中亦設計了「監理沙盒」（Regulatory Sandbox）機制，作為鼓勵創新與監管合作之重要工具。監理沙盒是一種監管工具，允許企業在監管機關的監督下，於限定期間內測試與實驗新穎且具創新性的產品、服務或商業模式。因此，監理沙盒具有雙重角色，其一為促進企業學習，即讓創新在真實世界環境中得以發展與測試；其二為支持監管學習，即制定實驗性的法律制度，協助並引導企業在創新過程中遵循監管要求[14]。

實務上，這種做法旨在於可控風險與監管機制之下推動實驗性創新，同時提升監管機關對新興科技的理解。歐盟就人工智慧監理沙盒的規定，主要在其人工智慧法（歐盟官方公報 2024 年 7 月 12 日公布版本）[15]中的第 57 條至第 62 條，條文規定歐盟成員國應設立一個或多個監理沙盒，以促進創新，特別是中小企業與新創企業，允許其在受控環境下測試高風險 AI 系統，並在符合 AI 法案規定下投入市場。

[14] 前述有關監理沙盒的定義及敘述，參見：European Parliament，Artificial intelligence act and regulatory sandboxes，網址：https://www.europarl.europa.eu/RegData/etudes/BRIE/2022/733544/EPRS_BRI(2022)733544_EN.pdf，最後瀏覽日期：2025 年 3 月 24 日。

[15] 歐盟人工智慧法的原始草案（2021 年版）已提及建立監理沙盒的概念，但內容較為粗略。於 2023 年 12 月協議草案中，增補了更具體的沙盒實施細節，包括：明確納入中小企業優先參與條款、擴充主管機關義務，包含提供指導文件、協助風險評估及提出跨境協調機制。顯示歐盟重視創新與嚴格監理之間的平衡；於歐盟官方公報 2024 年 7 月 12 日公布版本，將相關規定列於第 57 條至第 62 條。

（三）新加坡

新加坡 AI 政策發展主要由新加坡資訊通信媒體發展局（Infocomm Media Development Authority, IMDA）與個人資料保護委員會（Personal Data Protection Commission, PDPC）等部門主導，相關監管措施如下：

1. 人工智慧模型治理與管理框架[16]

2019 年 1 月 23 日 PDPC 發佈第 1 版「人工智慧模型治理與管理框架」(Model AI Governance Framework)，此為亞太地區首個針對私營部門組織實務應用而設計的 AI 倫理準則，以徵求更廣泛的諮詢、採納和回饋。該模型框架為私營部門組織提供了詳細且易於實施的指導，以解決部署人工智慧解決方案時的關鍵道德和治理問題。透過解釋人工智慧系統的工作原理、建立良好的資料問責實踐以及創建開放透明的溝通，該模型框架旨在促進公眾對技術的理解和信任。

2. 中小企業生成式 AI 沙盒[17]

新加坡企業發展局(Enterprise Singapore，ESG)和 IMDA 於 2024 年 2 月推出中小企業生成式 AI（GenAI）沙盒，該沙盒提供：(1)行銷與銷售類型之沙盒，可生成獨特的行銷內容和優化行銷策略等；(2)客戶互動類型之沙盒，讓中小企業可

[16] PDPC，Singapore’s Approach to AI Governance，網址：https://www.pdpc.gov.sg/help-and-resources/2020/01/model-ai-governance-framework，最後瀏覽日期：2025 年 3 月 20 日。

[17] IMDA，Singapore’s first generative AI Sandbox to familiarise and help SMEs get head start in capturing new AI opportunities，網址，https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/press-releases/2024/sg-first-genai-sandbox-for-smes?utm_source=chatgpt.com，2024 年 2 月 7 日，最後瀏覽日期：2025 年 6 月 5 日。

使用聊天機器人與客戶互動，為客戶生成客製化的建議及推薦資訊，以及自動執行客戶預訂和購買之需求等。通過申請之中小企業可試用其所選擇之解決方案進行為期 3 個月之試驗，試驗結束後，新加坡政府將蒐集中小企業的回饋，以評估這些解決方案的適用性與擴大採用的可行性。

除前述由政府主導的監管措施外，新加坡亦積極推動自願性的驗證機制。PDPC 發布的人工智慧模型治理與管理框架，雖為企業提供開發與部署 AI 系統時應考慮的原則，但在實務上，企業反映難以評估自身 AI 系統是否真正「可信任」、缺乏技術工具支援進行內部風險評估或對外說明，及缺乏標準化的評估基準來面對國際法規挑戰（如歐盟 AI 法案）。爰此，新加坡 IMDA 與 PDPC 於 2022 年共同推出「AI 驗證框架」（AI Verify），作為落實上述指引的工具。AI Verify 是一個自願性人工智慧系統驗證工具，幫助企業驗證其 AI 系統是否符合透明、公平、責任、安全等「可信任 AI」的核心原則。開發人員和所有者可以透過標準化測試，根據一系列原則驗證其人工智慧系統聲稱的性能[18]。

由於新加坡是一個小型開放的經濟體，在中美科技競爭與 AI 監管模式分歧（歐盟偏嚴、美國偏鬆）之下，若能提供中立、靈活且具國際適應力的 AI 治理工具，將有機會吸引跨國企業於該國部署與測試 AI 產品、強化與亞太及歐洲市場的互通標準及為區域制定 AI 技術規範建立領導地位。

[18] IMDA，Singapore launches world’s first AI testing framework and toolkit to promote transparency; Invites companies to pilot and contribute to international standards development，2022 年 5 月 25 日，網址：https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/press-releases/2022/sg-launches-worlds-first-ai-testing-framework-and-toolkit-to-promote-transparency，最後瀏覽日期：2025 年 3 月 25 日。

新加坡的人工智慧監管體系展現出高度靈活性與創新導向，透過監理沙盒機制與技術工具化治理方式，在創新與規範之間取得初步平衡。另新加坡推動的 AI Verify 驗證機制，不僅提供了新加坡國內治理需要，更是數位外交與技術標準戰略的一環。新加坡政府更進一步成立 AI Verify 基金會（Foundation），利用全球開源社群的集體力量和貢獻來開發 AI 測試工具，以負責任地使用 AI[19]，顯示其希望以開放協作模式擴大全球影響力。

二、有關 AI 治理之國際規範及多邊倡議

（一）聯合國「人工智慧倫理建議書」[20]

聯合國教育科學及文化組織（United Nations Educational, Scientific and Cultural Organization，UNESCO）在 2021 年 11 月 25 日宣布，193 個成員國一致通過一項人工智慧倫理全球協議「人工智慧倫理建議書」(The Recommendation on The Ethics of Artificial Intelligence)。該建議書旨在實現人工智慧給社會帶來的優勢並降低其帶來的風險，確保數位化轉型能尊重、保護和促進人權，解決透明度、問責機制和隱私方面的問題，並指出人工智慧正在加劇原已分裂和不平等的世界，應該加以控制。其核心原則包括：尊重人權與人類尊嚴、維護環境與永續發展、促進和平與社會正義、公平取用技術及性別平等與文化多元性。

---

[19] IMDA，Singapore launches AI Verify Foundation to shape the future of international AI standards through collaboration，2023 年 6 月 7 日，網址：https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/press-releases/2023/singapore-launches-ai-verify-foundation，最後瀏覽日期：2025 年 3 月 25 日。

[20] UNESCO，UNESCO member states adopt the first ever global agreement on the Ethics of Artificial Intelligence，2021 年 11 月 25 日，網址：https://www.unesco.org/en/articles/unesco-member-states-adopt-first-ever-global-agreement-ethics-artificial-intelligence，最後瀏覽日期：2025 年 3 月 25 日。建議書全文請見：聯合國教育科學及文化組織網頁，網址：https://unesdoc.unesco.org/ark:/48223/pf0000373434，最後瀏覽日期：2025 年 3 月 25 日。

（二）OECD 人工智慧建議書[21]

經濟合作暨發展組織（下稱經合組織）的人工智慧建議書（The OECD AI Principles）是國際間第一個關於人工智慧的政府間標準，於 2019 年通過，並於 2024 年更新，旨在促進人工智慧的負責任發展及應用，並為各國政府、企業和組織提供指導方針。該原則已被經合組織成員國及一些合作夥伴援引作為制定政策並創建人工智慧風險框架之參考，提倡創新、值得信賴、尊重人權和民主價值的人工智慧，為不同司法管轄區之間的全球互通性奠定一定基礎。其核心原則包括：以人為本與公平性（Inclusive and Human-centred Values）、透明性與可解釋性（Transparency and Explainability）、穩健性、安全性與可靠性（Robustness, Security and Safety）、問責性（Accountability）及永續發展與人類福祉（Sustainable Development and Well-being）。

OECD 人工智慧原則為人工智慧的負責任發展和應用提供了重要的指導方針，確保人工智慧在促進經濟社會發展的同時，也能夠尊重人權、保障安全和實現永續發展。

（三）ISO/IEC 42001[22]

AI 能夠以人類無法企及的速度和規模分析大量數據，識別模式和異常。隨著技術的進步和取得 AI 的便捷性增加，如何確保 AI 系統適用於其預定用途，並且使用得當，成為一大挑戰，並引發了 AI 的信任危機，因此 ISO/IEC 42001 - 人工智慧管理系統應運而生。ISO/IEC 42001 是全球第一個人工智慧（AI）管理系統的國際標準

[21] OECD，AI principles，網址：https://www.oecd.org/en/topics/ai-principles.html，最後瀏覽日期：2025 年 3 月 5 日。

[22] Bsi，ISO/IEC 42001 人工智慧管理－塑造值得信賴的人工智慧 (AI) 基礎，網址：https://www.bsigroup.com/zh-TW/standards/isoiec-42001/，最後瀏覽日期：2025 年 3 月 14 日。

，其制定背景主要源於企業組織需要規範指引，及建立大眾對 AI 的信任。按許多企業組織在應用 AI 技術時，面臨著諸如非透明自動決策、機器學習應用以及持續學習等問題，這些問題可能導致 AI 技術的使用存在風險；此外，為建立大眾對 AI 的信任，制定國際標準是縮小「AI 信心差距」並建立大眾對 AI 技術信任的關鍵。

ISO/IEC 42001 提出了一個「可認證」的人工智慧管理系統(AIMS)，實施這個標準，意味著使用「計畫-執行-檢查-行動」方法來制定與人工智慧相關的組織的健全治理政策和程序。標準並沒有太多詳述人工智慧的技術細節，而是提供了一個管理風險和機會的方法。這有助於最大限度地減少與人工智慧相關的潛在風險和負面影響。

（四）國際間之多邊倡議

全球各大多邊組織與國際高峰會（如 G7、G20）及其他 AI 相關國際論壇，均已將「可信任 AI（Trustworthy AI）」納入討論與政策協議的核心之一。以下整理近年來 G7、G20、及各重要 AI 國際高峰會提出的主要宣言、協議、建議與規範指引之內容與重點：

1. 2023 年 G7 廣島 AI 進程（Hiroshima AI Process）聲明[23]

隨著人工智慧快速普及並深入政府、企業與社會日常，國際社會逐漸認知到，新興數位科技的治理未跟上科技發展的腳步，國際間缺乏共識，導致全球 AI 治理出現碎片化與標準不一的問題，在缺乏規範的狀況下，AI 可能侵犯隱私、擴大偏見、傳播錯誤訊息或被用於操控輿論。因此，G7

[23] 聲明全文參見：2023 年廣島 G7 峰會領導人聯合聲明（G7 Hiroshima Leaders' Communiqué），日本外務省網頁，2023 年 5 月 20 日，網址：https://www.mofa.go.jp/policy/economy/summit/hiroshima23/documents/pdf/Leaders_Communique_01_en.pdf，最後瀏覽日期：2025 年 3 月 20 日。其中涉及人工智慧部分，在該聲明第 38 段。

（七大工業國組織高峰會議）將 AI 治理作為優先議題，推動以民主價值為核心的可信任 AI。2023 年廣島 AI 進程聲明提及，數位經濟的治理應持續依據成員共同的民主價值進行更新，包括：公平性、問責制、透明性、安全性、防範線上騷擾、仇恨與濫用、尊重隱私與人權、基本自由及個人資料保護。並承諾進一步推進 AI 標準的多方參與訂定，及尊重法律框架，強調建立促進負責任 AI 的程序。

G7 強調國際間對 AI 治理與治理框架互通性的討論重要性，同時承認為實現「可信任的 AI」之共同願景與目標，G7 各成員國可能採用不同的方法與政策工具。支持透過多方利害關係人組成的國際組織發展可信任 AI 的工具，並鼓勵在標準制定機構中採取多方參與的過程以推動國際技術標準的制定。鼓勵 OECD 等國際組織分析政策發展的影響，也鼓勵全球 AI 夥伴關係（GPAI）執行實際專案。指示相關部長透過 G7 工作小組、在 OECD 和 GPAI 的協助下，以包容的方式建立「廣島 AI 進程」，就生成式 AI 展開討論。這些討論可包括：治理、智慧財產權（包括版權）的保護、透明度的促進、對外國資訊操控（包括假訊息）的因應，及技術的負責任使用。

2.2023 年 G20 印度峰會領袖宣言[24]

2023 年 G20 印度峰會領袖宣言第 61 節提及負責任的運用人工智慧（AI）造福全民。AI 的迅速進展為全球數位經

[24] 聲明全文參見：2023 年 G20 印度峰會領袖宣言（G20 New Delhi Leaders' Declaration），印度政府外交部網頁，2023 年 9 月 9 日-10 日，網址：https://www.mea.gov.in/Images/CPV/G20-New-Delhi-Leaders-Declaration.pdf，最後瀏覽日期：2025 年 3 月 20 日。

濟帶來繁榮與擴張的前景。各國努力以包容、以人為本、負責任的方式善用 AI 解決全球挑戰，同時保障人民權利與安全。為確保 AI 的負責任發展、部署與使用，必須處理以下關鍵議題：人權保護、透明與可解釋性、公平與問責、合理監管與安全、人為監督、倫理風險與偏誤、隱私與資料保護（第 57 至第 61 節專章討論數位經濟與人工智慧），並提出建立全球 AI 治理架構之必要性。

3. 2025 年法國人工智慧行動高峰會聯合聲明[25]

2025 年 2 月在法國舉辦的人工智慧行動高峰會中，宣布了全球 AI 聯合聲明，呼籲一個「開放」、「包容」和「道德」的 AI 發展，承諾共同強化需要「全球對話」的 AI 治理，避免「市場集中」，使全球人民更容易取得這項科技，並於高峰會上宣布正式成立由國際能源總署（IEA）領導的 AI 能源衝擊監察機構，以及聚集相關產業主要企業的永續 AI 聯盟，持續推進 AI 的國際治理。

惟美國和英國卻拒絕簽署前揭聲明，美國在高峰會中表示，將保持美國在 AI 產業上的領導地位，同時警告對 AI 的「過度監管」，恐將「扼殺由美國主導、正在蓬勃發展的產業」。英國則表示，英國因「國家利益」而沒有簽署[26]。美國與英國的拒絕簽署，凸顯了在該議題上兩個價值觀的衝突，

[25] 聲明全文，ELYSEE，Statement on Inclusive and Sustainable Artificial Intelligence for People and the Planet.，2025 年 2 月 11 日，網址：https://www.elysee.fr/en/emmanuel-macron/2025/02/11/statement-on-inclusive-and-sustainable-artificial-intelligence-for-people-and-the-planet，最後瀏覽日期：2025 年 3 月 18 日。

[26] 曾婷瑄，美國反過度監管拒簽 AI 峰會宣言 稱不與獨裁政權合作，中央通訊社，2025 年 2 月 12 日，網址：https://www.cna.com.tw/news/aopl/202502120009.aspx，最後瀏覽日期：2025 年 3 月 18 日。

而各國於 AI 治理上出現的分歧，尤需經由國際合作解決[27]。

三、外國立法例及國際規範治理機制之借鏡分析

AI 所帶來的創新效益固然顯著，然其伴隨之風險與倫理挑戰亦逐漸浮現，對於個人權益、資料保護、社會公平與國家安全構成深遠影響。面對此一發展趨勢，世界主要國家與國際組織相繼提出「可信任 AI」(Trustworthy AI) 之治理架構與立法行動，試圖在促進創新與保障公共利益間取得平衡。美國透過行政命令與風險管理框架強化責任機制；歐盟則以人工智慧法 (AI Act) 建立用途與風險導向之全面規範；新加坡採取監理沙盒與政策指引相結合的柔性治理；UNESCO、OECD、G7 等國際組織亦提出共同原則，推動跨國合作與標準一致性。

前述外國立法例及國際規範中，可資我國在人工智慧立法上借鏡之處有，風險分類、人類監督、偏誤防範、透明揭露、隱私保障、監理沙盒制度等，茲彙整如表 2：

表 2：AI 治理機制外國立法例彙整表

法域/組織	核心治理策略	借鏡價值	潛在風險/限制
美國	NIST AI 風險管理框架、總統行政命令；多方參與、自律導向	彈性高、鼓勵創新；強調透明性與風險控制工具	缺乏統一聯邦法規，公私落差大
歐盟	制定《AI 法案》，風險導向、分類式監管；強制規範高風險應用。設	架構明確，合規流程清楚；人權保障、技術透明並重	法規複雜，合規成本高；中小企業承擔壓力大

[27] 李淑瓊，AI 跨國監管之問題研析，立法院法制局議題研析（編號：R02716），2025 年 3 月 25 日，網址：https://www.ly.gov.tw/Pages/Detail.aspx?nodeid=6590&pid=249383，最後瀏覽日期：2025 年 5 月 1 日。

	有監理沙盒制度。		
新加坡	Model AI Governance Framework;倫理導向＋實作建議；沙盒試驗	實用、操作性強；善用沙盒與實證機制	非強制，法制效力相對有限
UNESCO /OECD	倫理原則、透明治理、人權導向；推動國際標準一致性	提供價值與原則架構，便於全球對接	無強制效力，需轉化為國內立法支持

資料來源：作者彙整各國立法例及國際規範

肆、 我國推動可信任 AI 的政策及法規現況

我國政府為掌握 AI 發展的契機，宣示 2017 年為台灣 AI 元年後，於同年 8 月推出「AI 科研戰略」，並於 2018 年 1 月 18 日起推動 4 年期的「台灣 AI 行動計畫」(2018 年至 2021 年)，全面啟動產業 AI 化 。科技部[28]並於 2019 年 9 月訂定「人工智慧科研發展指引」(AI Technology R&D Guidelines)，強調「以人為本」、「永續發展」及「多元包容」三大核心價值，並衍生構築 8 大指引，包括「共榮共利」、「公平性與非歧視性」、「自主權與控制權」、「安全性」、「個人隱私與數據治理」、「透明性與可追溯性」、「可解釋性」及「問責與溝通」等。另外，各別主管機關亦有因應業務需要訂定 AI 相關指引者，例如：行政院訂定之「行政院及所屬機關（構）使用生成式 AI 參考指引」、金融監督管理委員會訂定之「金融業運用人工智慧(AI)指引」及「金融業運用人工智慧（AI）之核心原則與相關推動政策」，及衛生福利部針對 AI 醫療器材軟體查驗登記訂定之「人工智慧機器學習技術之醫療器材軟體查驗登記技術指引」。惟隨著科技的進步，前述指引已無法應付人工智慧的迅速發展，為了更能具體落實透明、可控制性及隱私保護等原則，國家科學及技術委員會（下稱國科會）已於 2024

[28] 2022 年 7 月 27 日配合行政院組織調整，改制為「國家科學及技術委員會」。

年 7 月 15 日預告制定「人工智慧基本法」草案[29]，該草案涵蓋 AI 法律名詞定義、隱私保護、風險管控、倫理原則等面向。

此外，數位發展部為建立符合國家與產業需求之人工智慧產品與系統評測機制，並健全國內 AI 評測制度之發展環境，成立「AI 產品與系統評測中心」(Artificial Intelligence Evaluation Center，AIEC)[30]。另數位發展部為協助各機關善用人工智慧技術，解決業務痛點，參酌國際間陸續頒布之 AI 相關指引、指南文件，研擬「公部門人工智慧應用參考手冊」(草案)，供各機關參考並建置 AI 服務，自 2025 年 1 月 1 日至 12 月 31 日進行為期一年之試辦作業[31]。值得注意的是，基於防範公務機關資安風險等考量，行政院於今（2025）年 2 月 20 日已宣布公務機關立即全面禁用由中國杭州深度求索人工智慧基礎技術研究有限公司所發布的 DeepSeek AI 服務，包含雲端服務、APP 及地端下載等方式，公務機關皆不得使用，惟目前未進一步限制業者提供民眾下載及使用，政府表示將持續關注 DeepSeek 議題，並提醒民眾要注意相關風險[32]。

伍、 問題研析與建議

一、我國推動可信任 AI 立法策略之比較與評估

如前所述全球面對人工智慧（AI）快速發展所帶來的倫理、法治與社會衝擊，已逐步展開相關治理制度之建構。綜觀主要國家與國際

[29] 國家科學及技術委員會，同註 1。

[30] AI 產品與系統評測中心，成立背景，網址：https://www.aiec.org.tw/AboutUs/EstablishmentBackground，最後瀏覽日期：2025 年 3 月 18 日。

[31] 數位發展部，公部門人工智慧應用參考手冊（草案），2025 年 1 月 8 日，網址：https://moda.gov.tw/digital-affairs/digital-service/guide/15002，最後瀏覽日期：2025 年 3 月 18 日。

[32] 數位發展部資通安全署，公務機關全面禁用 DeepSeek AI 服務，未限制一般民間使用，數位發展部資通安全署新聞稿，2025 年 2 月 20 日，網址：https://moda.gov.tw/ACS/press/news/press/15296，最後瀏覽日期：2025 年 3 月 21 日。

組織之立法與政策規範，可信任 AI 的治理模式可大致分為兩大類型：其一為以歐盟為代表的預防性、法規導向模式，強調透過強制立法規範風險分級、透明性與人權保障，以建構完整之合規架構；其二為以美國為代表的發展導向、自律管理模式，偏重產業自主與創新鼓勵，透過行政指導與技術標準引導市場自我調整。兩者在法制設計、監管角色、政策工具與市場介入程度上均有顯著差異，成為當前全球可信任 AI 治理架構發展的兩大典型模式，對其他國家在制度選擇上形成重要參考依據。茲將兩種治理模式比較分析如表 3：

（一）歐盟與美國 AI 監管模式比較

表 3：歐盟與美國 AI 監管模式比較表

比較面向	歐盟	美國
監管態度	謹慎、預防性	寬鬆、友善創新
立法架構	人工智慧法，分類風險等級，對高風險 AI 產品實施強制合法性要求	行政命令及行業指引，強調政府部門先行示範，無統一法規
重點機制	強制評估、資料治理、透明度要求、風險分類系統	鼓勵產業自律、創新沙盒、風險導向原則
核心理念	保障基本權利、公平、透明與問責	促進創新、維持產業競爭力、國家安全

資料來源：作者自製

（二）建議應先確定監管模式後再進行法規草擬

我國目前未有監管人工智慧的專法，倘未來要制定專法，在進行法律條文草擬階段前，應先確立整體監管策略之方向，即應先確定係採偏向歐盟模式的預防性嚴格監管，或美國模式的發展導向風險管理，或參考新加坡之發展模式，以確定法律設計邏輯、政府角色定位及

政策資源分配。

在進行決策時，需考量我國人工智慧產業發展的程度、主要的出口市場、政府執行監管的成本、公民社會對權利保障的意識、嚴格監管對企業創新的影響等，例如：我國是否已有成熟 AI 應用產業？產業能否承擔法律遵循的成本？社會對隱私、演算法歧視、深偽技術等議題的敏感程度等因素。

（三）針對我國就可信任 AI 監管模式之建議

我國 AI 產業尚處擴展期，主要應用於智慧製造、智慧醫療及金融科技，企業以中小企業為主，缺乏具全球競爭力的大型企業。當前全球 AI 市場高度集中，主要客戶群集中於技術先進國家如美國、歐盟、日本與新加坡，受限於高端技術需求，我國 AI 產品出口與市場拓展面臨限制。加上地緣政治與國安風險，對中國大陸的技術輸出更具挑戰，凸顯我國 AI 產品出口策略與監管模式之重要性。在此背景下，若我國過早採行如歐盟般嚴格的預防性監管，將提高產業法律遵循成本，削弱國際競爭力。中小企業如需負擔高額合規成本（如風險分級、外部審查、可解釋性等），將難以與國際大型企業競爭，恐抑制產業活力與創新。因此，建議我國在 AI 立法政策上可採取類似美國的發展導向與彈性自律模式，鼓勵創新並保留政策調整空間，以降低產業法律遵循負擔、提升產業反應速度與彈性。2024 年 7 月 15 日國科會預告之「人工智慧基本法」草案，即以產業發展為核心，在保障人權與倫理議題上採漸進、合作、自律的方式，顯示其監管模式傾向美國的寬鬆而非歐盟的嚴格監管架構。建議主管機關未來持續關注國際趨勢與社會對 AI 風險的關注程度，保持政策彈性，評估是否適度引入歐盟或新加坡模式中強化治理與創新平衡的元素。

二、強化可信任 AI 之驗證環境

（一）建置可信任 AI 驗證環境之必要性

所謂「可信任 AI」，乃指具備合法、合倫理、穩健可靠等特性的人工智慧系統。根據 OECD 與歐盟的相關原則，可信任 AI 應符合數據治理、演算法透明、可解釋性、公平性、問責性與安全性等多項指標。若 AI 系統欠缺這些基本保障，不僅可能產生偏誤歧視、資訊濫用或損害人權，更會削弱使用者與社會大眾對 AI 技術的信任，進而影響其應用普及與永續發展。惟如何去驗證各別 AI 系統是否符合前述可信任 AI 應具備之要件及指標，則有待第三方認證及驗證機構之驗證，有鑑於此，建立一套能提升 AI 系統透明性、公平性與可靠性的可信任 AI 驗證制度，即為推動 AI 治理的重要政策工具。

（二）我國現況與建議

數位發展部於 2024 年 3 月 18 日訂定發布「AI 產品與系統評測中心設置要點」，並成立了「AI 產品與系統評測中心」(AIEC)，以建立 AI 產品與系統的評測機制，提升 AI 應用的可信度與安全性，促進可信任 AI 環境發展。該中心依循行政院臺灣 AI 行動計畫 2.0（2023-2026)，在數位發展部數位產業署的支持下，國家資通安全研究院打造首座「AI 驗證機構」，與工業技術研究 AI 測試實驗室共同致力於協助國內 AI 產業朝安全可靠方向發展，目前 AI 驗證機構正處於試運行階段，在此期間係以評價服務取代正式驗證。通過提供專業的評價服務，為企業及民眾打造更為堅實的 AI 信任基礎。隨著試運行期間的經驗累積，未來將逐步完善驗證機制，以適應快速發展的 AI 技術，確保台灣 AI 產業的長期健康發展[33]。AI 產品與系統評測中心

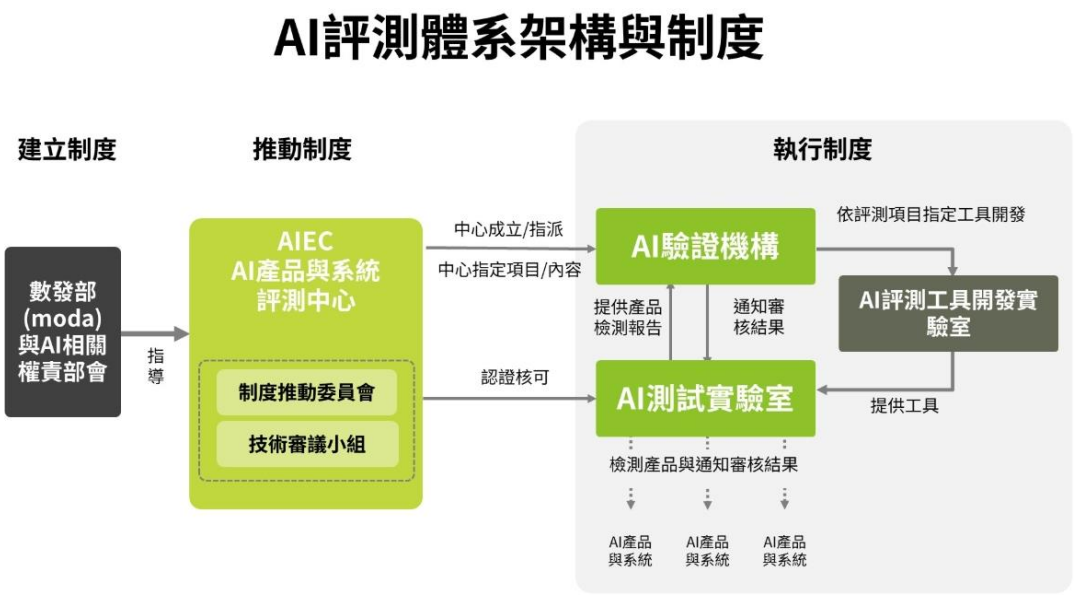
[33] AI 產品與系統評測中心，關於 AI 驗證機構，網址：https://www.aiec.org.tw/AICertificationBody/Index，最後瀏覽日期：2025 年 3 月 31 日。

將規劃並建立國內 AI 評測體系，並擬由以下三個機構組成[34]：

1. AI 評測中心：AI 評測制度與方法制定。

2. AI 驗證機構：AI 評測制度之執行單位，負責產出評價報告。

3. AI 測試實驗室：AI 評測中心認可之測試報告產出者。



資料來源：AI 產品與系統評測中心[35]

AI 產品與系統評測中心係依據「AI 產品與系統評測中心設置要點」設立，該要點僅為行政規則，目前作法缺乏法源依據。惟「人工智慧基本法」草案第 9 條已明定，政府應避免 AI 應用危害國民安全與社會秩序，並賦予數位發展部及其他相關機關得提供或建議「評估驗證」工具或方法之權限，提供驗證業務具體法源依據，較為周延。

AI 產品與系統評測中心擬打造之「AI 驗證機構」旨在為國內 AI 產業提供評價服務，降低 AI 產品帶來的風險。其首要任務是針對 AI

[34] AI 產品與系統評測中心，同註 30。
[35] 同前註。

產品進行評價，確保 AI 產品之風險得到有效控管，並提升大眾對 AI 技術的信賴。AI 評測服務中，驗證機構將對測試實驗室產出之測試報告進行評價，並進一步出具評價報告對測試結果進行相關說明[36]。

前述 AI 評測體系係著重於為國內 AI 產業提供評價（驗證）服務[37]，著重於建立國內的評測制度與標準，強調與國際接軌，並提供企業專業的評測服務。而新加坡的 AI Verify 則更偏向於推動全球 AI 治理的標準化，透過開源社群的合作，促進 AI 測試框架與最佳實踐的發展，值得臺灣借鏡。AIEC 可考慮引入開源工具與框架，提升測試的透明度與可驗證性，並加強與國際社群的連結，提升臺灣在全球 AI 治理領域的影響力。

三、建置 AI 監理沙盒機制以促進創新

當前人工智慧技術快速發展，應用領域不斷擴展，為我國產業帶來創新動能的同時，也引發如資料保護、演算法歧視、責任歸屬等多重法律與倫理風險。有鑒於此，建立一套兼顧創新與風險控管的法制架構，已成為國家科技治理的迫切課題。透過監理沙盒制度，可在受控環境中測試新技術，有效管控其倫理、法律與社會影響。其次，由於 AI 發展快速，既有法律可能無法即時適應新型 AI 應用，透過沙盒測試與實務回饋，可作為法律與政策調整的依據。是以，監理沙盒制度的建置可兼顧創新與風險控管，並可提升政策靈活性與適應性。

面對 AI 治理高度不確定性與技術變動速度快的特性，建議我國可參考歐盟與新加坡等國際先進法制經驗，建置具彈性、實驗性質的

[36] AI 產品與系統評測中心，同註 30。

[37] 根據媒體報導「數產署表示，AI 評測中心目前已接獲多國內廠商來送測，今年可望逾十家業者送驗，包括聯發科（2454）、長春石化（1764）、台智雲（7735）等。」。余弦妙，數發部為強化我國 AI 建置 AIEC 今年可望評測逾十家台廠，聯合新聞網，2025 年 2 月 6 日，網址：https://udn.com/news/story/7240/8530997，最後瀏覽日期：2025 年 3 月 18 日。

AI 監理沙盒機制，前揭草案第 6 條之立法說明雖有提及，惟條文內容過於抽象籠統，建議於該條增訂第 2 項，規定各目的事業主管機關為建立或完備創新實驗環境，得提供特定範圍、期間及對象之法規適用彈性，作為我國 AI 法制化進程的基礎工程。爰以國科會預告之「人工智慧基本法」草案第 6 條為基礎，提出增訂第 2 項之立法建議如下：「各目的事業主管機關為建立或完備創新實驗環境，得提供特定範圍、期間及對象之法規適用彈性，允許人工智慧創新產品或服務在符合法定條件下試行，藉以觀察其可行性、風險影響及監理需求。」。

四、發展主權 AI 以支持可信任的 AI

「主權 AI」是指由國家主導或支持，基於國家戰略需求，自主研發、建置及運用之人工智慧系統。亦即，一個國家或地區擁有自主開發、控制和運用人工智慧技術的能力，包括資料的收集、模型的訓練、以及技術的應用等，以確保國家安全、經濟發展和文化自主性。

（一）建立主權 AI 的必要性

為確保 AI 技術與應用不受外部掌控、強化國安與數位主權，我國亟需推動主權 AI 法制化。首先，國家安全是主權 AI 發展最核心的戰略需求，現代戰爭中 AI 已是決策與情資判讀的關鍵技術，若關鍵系統仰賴國外平台，恐導致資安脆弱、受制技術封鎖。有學者指出，AI 作為一種工具，可以是國家用來強化自身安全的利器，亦可以是弱化他國安全之兵器。以 AI 操控認知作戰為例，其帶來之威脅不容小覷，例如，俄烏戰爭期間，莫斯科透過社群媒體及網路輿情操作，以 AI 大量投送具特定目的訊息，試圖影響國際社會的觀點和看法，並透過操縱訊息，支持其在烏克蘭之地緣政治目標。此一操作，不

僅削弱了民主政治，更對烏克蘭主權和安全構成威脅[38]。

其次，公共服務日益依賴 AI 決策輔助，若系統架構與語言資料訓練依賴國外平台，恐造成資料主權流失與民眾隱私風險。制定資料治理標準、採購規範，能強化安全、公平的公共 AI 應用。中國大陸近年來積極發展自己的 AI 技術，推出如 DeepSeek、文心一言等大型語言模型，當我國的使用者開始大量使用這類產品時，其不僅可以了解我方用戶的思維模式，還可透過 AI 技術影響我方的社會文化與價值觀[39]。當務之急，積極發展本土語料與語言習慣訓練之大型語言模型，以強化在地溝通及語言文化傳承，維護我方數位主權及民主韌性。

最後，關鍵基礎設施如能源、交通與金融系統等類 AI 模組，若仰賴外部封閉模型，將削弱系統可控性與安全性。為確保系統可控性、可審計性與應變能力，政府應透過立法要求關鍵基礎設施 AI 採用主權模型或具備透明性、安全性之系統，並建立跨部會監管架構。

（二）打造「可信任 AI 對話引擎」

我國已與輝達等世界頂尖 AI 晶片巨擘合作，共同打造 AI 資料中心及超級電腦，用於訓練我國的「可信任 AI 對話引擎」（Trustworthy AI Dialogue Engine, TAIDE）。TAIDE 是一種大型生成式 AI 語言模型，採用臉書的母公司 Meta 的開源語言模型 Llama 3，適用我國的繁體中文。從主權 AI 的層面來看，

[38] 趙萃文，〈人工智慧對國家安全的影響與因應〉，《安全與情報研究》，第 7 卷，第 2 期，2024 年 7 月，頁 47、51。

[39] 廖明輝，自由開講》主權 AI 崛起 TAIDE 捍衛數位主權，自由時報，2024 年 9 月 25 日，網址：https://talk.ltn.com.tw/article/breakingnews/4809861，最後瀏覽日期：2025 年 3 月 28 日。

擁有屬於我國的正體中文大型語言模型對於未來的技術發展非常重要。這樣的語言模型將有助於我國更好地掌握 AI 技術，並減少對其他國家技術的依賴[40]。

（三）發展主權 AI 的法制建議

主權 AI 的建立不只是科技競爭問題，更是國家戰略層級的挑戰，我國應及早布局主權 AI 系統與可信任 AI 機制，避免在全球 AI 治理中被動應對。透過整合政府資源、強化本地語言模型、建構基礎設施、開放合法語料及培育人才，臺灣有機會建立兼具創新與信任的 AI 典範，並在全球數位競爭中維持主體性與領導力。我國雖已投入預算，用於 AI 資料中心和相關發展，發展可信任 AI 對話引擎（TAIDE），惟相關政策之推動欠缺法律依據，前述國科會預告之「人工智慧基本法」草案亦欠缺相關規定，爰以預告該草案內容為基礎，提出相關條文立法建議之條文對照表如下，以建構推動主權 AI 之法律架構，確保其正當性與持續性：

國科會「人工智慧基本法」預告草案條文對照表

預告草案條文	本報告建議條文	說明
第一條 為促進以人為本之人工智慧研發與應用，維護國民生命、身體、健康、安全及權利，提升國民生活福祉、維護國家文化價值及	第一條 為促進以人為本之人工智慧研發與應用，維護國民生命、身體、健康、安全及權利，提升國民生活福祉、維護國家文化價值及	**草案說明：** 一、本法之立法目的。 二、人工智慧為攸關國家發展之戰略性科技，為積極發展與應用人工智慧，強化 與深耕

[40] 李先泰，強化主權 AI！國科會證實「3 年內砸逾 900 億元」，為何台灣不惜血本也要升級算力？，數位時代，2024 年 11 月 21 日，網址：https://www.bnext.com.tw/article/81378/taiwan-ai-power-up?，最後瀏覽日期：2025 年 3 月 28 日。

國家競爭力，增進社會國家之永續發展，特制定本法。	確保國家在人工智慧領域之自主性、安全性及競爭力，增進社會國家之永續發展，特制定本法。	以人為本之人工智慧技術，促進技術應用與產業發展，同時維護人民安全、國家安全及保障人民權利，以期人工智慧可回應人文與社會發展所需，邁向社會永續發展。因此，發展與應用人 工智慧之同時，有賴於制定具有指標與引導性原則之立法，以作為發展人工智慧之規範與促進應用之法源基礎。 **本報告說明：** 國科會預告草案之立法目的欠缺發展主權人工智慧之規定，爰增訂「確保國家在人工智慧領域之自主性、安全性」等文字，以確立我國推動主權人工智慧之法律架構，確保其正當性與持續性。
第二條 本法所稱人工智慧，係指以機器為基礎之系統，該系統具自主運行能力，透過輸入或感測，經由機器學習與演算法，可為明確或隱含之目標實現預測、內容、建議或決策等影響實體或虛擬環境之產	第二條 本法所稱人工智慧，係指以機器為基礎之系統，該系統具自主運行能力，透過輸入或感測，經由機器學習與演算法，可為明確或隱含之目標實現預測、內容、建議或決策等影響實體或虛擬環境之產	**草案說明：** 參考美國國家人工智慧創新法案（National AI Initiative Act of 二〇二〇）美國法典(U.S. Code)第九四〇一章、國際標準化組織（ISO）及國際電工委員會（IEC）聯合制定技術規範

出。	出。 本法所稱國家主權人工智慧，係指由國家主導或支持，基於國家戰略需求，自主研發、建置及運用之人工智慧系統。	(ISO/IEC)四二○○一：二○二三人工智慧管理系統、美國國家標準暨技術研究院(National Institute of Standards and Technology，NIST)AI風險管理框架，以及歐盟人工智慧法(Artificial Intelligence Act)對於人工智慧系統之定義，說明人工智慧必須被設計為具備一定程度之自主運行能力，透過輸入（input）或感測（sensing），可為明確（explicit）或隱含（implicit）之特定目的（objectives），經過機器學習（machine-learning）與演算法（algorithms）實現諸如預測、內容、建議或決策(such as predictions, content, recommendations, or decisions)等影響實體或虛擬環境之產出，與其他軟體系統有別。 **本報告說明：** 由於國科會預告草案欠缺發展主權人工智慧之定義，爰於第二項增訂「國家主權人工智慧」之定義規定。

第四條 政府應積極推動人工智慧研發、應用及基礎建設，妥善規劃資源整體配置，並辦理人工智慧相關產業之補助、委託、出資、獎勵、輔導，或提供租稅、金融等財政優惠措施。	第四條 政府應積極推動人工智慧研發、應用及基礎建設。 政府應致力建構國家主權人工智慧，以確保國家安全、公共服務、關鍵基礎設施、語言文化傳承及重要產業之自主控制能力。 政府應妥善規劃資源整體配置，並辦理人工智慧相關產業之補助、委託、出資、獎勵、輔導，或提供租稅、金融等財政優惠措施。	**草案說明：** 人工智慧發展與應用涉及領域甚廣，其整體資源規劃，應由政府各機關依其業務職掌負責辦理，爰參考產業創新條例第九條及科學技術基本法第六條，定明政府機關應推動人工智慧發展等運用之方式。 **本報告說明：** 國科會預告草案未明定發展主權人工智慧之內涵，爰於第二項增訂政府應建構國家主權人工智慧之具體規定，並將原第一項後段規定移列至第三項。
第六條 為促進人工智慧技術創新與永續發展，各目的事業主管機關得針對人工智慧創新產品或服務，建立或完備既有人工智慧研發與應用服務之創新實驗環境。	第六條 為促進人工智慧技術創新與永續發展，各目的事業主管機關得針對人工智慧創新產品或服務，建立或完備既有人工智慧研發與應用服務之創新實驗環境。 各目的事業主管機關為建立或完備創新實驗環境，得提供特定範圍、期間及對象之法規適用彈性，允許人工智慧創新產品或服務在符合法定條件下	**草案說明：** 參考歐盟人工智慧法，鼓勵其會員國政府建立人工智慧實驗沙盒制度（Regulatory Sandbox），提供一個受控環境，以促進人工智慧之創新，使其於投放市場或入使用之前，可於有限時間開發、測試和驗證。爰定明各目的事業主管機關應建立或完備有關人工智慧研發與應用之創新實驗環境，進一步使國民受益於人工智慧創新科技。

	試行，藉以觀察其可行性、風險影響及監理需求。	**本報告說明：** 面對人工智慧治理高度不確定性與技術變動速度快的特性，建議我國可參考歐盟與新加坡之制度，建置具彈性、實驗性質的 AI 監理沙盒機制，本條預告草案之立法說明雖有提及，惟條文內容過於抽象籠統，建議增訂各目的事業主管機關為建立或完備創新實驗環境，得提供特定範圍、期間及對象之法規適用彈性規定，作為我國 AI 法制化進程的基礎工程。
第十五條 政府應建立資料開放、共享與再利用機制，提升人工智慧使用資料之可利用性，並定期檢視與調整相關法令及規範 。 政府應致力提升我國人工智慧使用資料之品質與數量，確保訓練結果維繫國家多元文化價值與維護智慧財產權。	第十五條 政府應建立資料開放、共享與再利用機制，提升人工智慧使用資料之可利用性，並定期檢視與調整相關法令及規範 。 政府應致力提升我國人工智慧使用資料之品質與數量，蒐集本國語言、文化、法律等相關資料，提供人工智慧模型訓練運用，確保訓練結果維繫國家多元文化價值與維護智慧財產權。	**草案說明：** 一、資料為人工智慧發展之重要元素，政府有必要確保人工智慧之創新與產業發得以取高品質和可追溯的資料。參考歐盟人工智慧法有關支援高品質資料近用之規定，以及美國 AI 行政命令要求聯邦資料長委員會研擬開放相關指引等規定。爰於第一項定明政府需就相關規範定期檢視並為必要調整，俾利人工智慧

		發展所需。 二、為確保 AI 訓練結果維繫國家文化價值，避免影響弱勢、多元族群權益及人民之智慧財產權，於第二項定明各機關應致力推動之事項，以完善我國資料治理機制。 **本報告說明：** 為發展主權人工智慧，維護我國資料與語言之主體性，並提供主權 AI 訓練運用，爰增訂第二項。

陸、 結論

在人工智慧技術快速發展，應用範圍日益擴大的今日，如何於促進創新與保障公共利益之間取得平衡，成為各國 AI 政策與治理機制設計的核心課題。美國強調風險導向與產業自律，歐盟則透過人工智慧法建立高標準之法律規範，新加坡則具實用及操作性。我國 AI 產業尚處發展擴張期，且以中小企業為主，面對全球市場與技術快速演進的競爭壓力，建議短期內應採取較具彈性之監管模式，以鼓勵創新為主，並降低企業之法律遵行負擔。惟在採取彈性監管模式的同時，仍應建構可信任的 AI，例如：建置可信任 AI 之驗證環境、建立 AI 監理沙盒機制、強化資料治理與技術透明等，以回應社會對 AI 風險的關切。此外，為因應數位主權與語言文化保護的戰略需求，推動「主權 AI」的發展亦勢在必行，透過發展繁體中文大型語言模型如

TAIDE，確保我國在 AI 領域的自主性、安全性與文化延續性，並作為建構可信任的 AI 之基礎。

參考文獻

一、期刊論文

1、趙萃文，〈人工智慧對國家安全的影響與因應〉，《安全與情報研究》，第 7 卷，第 2 期，2024 年 7 月，頁 43-84。

2、Kaminski, Margot E.，〈The Developing Law of AI: A Turn to Risk Regulation. The Digital Social Contract: A Lawfare Paper Series〉，《U of Colorado Law Legal Studies Research Paper》，No. 24-5，2023 年 4 月 2 日，網址：https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4692562#，最後瀏覽日期：2025 年 5 月 1 日。

二、網路資源

1、余弦妙，數發部為強化我國 AI 建置 AIEC 今年可望評測逾十家台廠，聯合新聞網，2025 年 2 月 6 日，網址：https://udn.com/news/story/7240/8530997，最後瀏覽日期：2025 年 3 月 18 日。

2、李淑瓊，AI 跨國監管之問題研析，立法院法制局議題研析（編號：R02716），2025 年 3 月 25 日，網址：https://www.ly.gov.tw/Pages/Detail.aspx?nodeid=6590&pid=249383，最後瀏覽日期：2025 年 5 月 1 日。

3、李先泰，強化主權 AI！國科會證實「3 年內砸逾 900 億元」，為何台灣不惜血本也要升級算力？，數位時代，2024 年 11 月 21 日，網址：https://www.bnext.com.tw/article/81378/taiwan-ai-power-up?，最後瀏覽日期：2025 年 3 月 28 日。

4、陳箴，美國國家標準與技術研究院公布人工智慧風險管理框架（A

I RMF 1.0)，財團法人資訊工業策進會科技法律研究所，2023 年 4 月，網址：https://stli.iii.org.tw//article-detail.aspx?tp=1&d=8974&no=64，最後瀏覽日期：2025 年 3 月 28 日。

5、陳岩，人工智能大爆發後的監管命題：風險分類、路線分化與地緣政治陰影，BBCNEWS 中文，2025 年 3 月 14 日，網址：https://www.bbc.com/zhongwen/articles/ckgnvl7pwr3o/trad?at_link_id=252AB14A-02CA-11F0-BAA9-BE9226ED606D&at_campaign=Social_Flow&at_link_origin=bbc_news_trad&at_medium=social&at_campaign_type=owned&at_ptr_name=facebook_page&at_format=link&at_bbc_team=editorial&at_link_type=web_link&fbclid=IwY2xjawJEheVleHRuA2FlbQIxMQABHS6upP0vvfzn0yIkLSZswxUkz6cSZcfxQSsAORUwb3Gp93MNirmw01861Q_aem_mNscyVHf5qjjbf3X5zfoag，最後瀏覽日期：2025 年 3 月 17 日。

6、國家科學及技術委員會，國家科學及技術委員會公告：預告制定「人工智慧基本法」草案，公共政策網路參與平臺，2024 年 7 月 15 日，網址：https://join.gov.tw/policies/detail/4c714d85-ab9f-4b17-8335-f13b31148dc4，最後瀏覽日期：2025 年 3 月 14 日。

7、曾婷瑄，美國反過度監管拒簽 AI 峰會宣言 稱不與獨裁政權合作，中央通訊社，2025 年 2 月 12 日，網址：https://www.cna.com.tw/news/aopl/202502120009.aspx，最後瀏覽日期：2025 年 3 月 5 日。

8、臺灣 AI 卓越中心，美國白宮發布 AI 行政命令，網址：https://www.twaicoe.org/%E7%BE%8E%E5%9C%8B%E7%99%BD%E5%AE%AE%E7%99%BC%E5%B8%83AI%E8%A1%8C%E6%94%BF%E5%91%BD%E4%BB%A4，最後瀏覽日期：2025 年 3 月 5 日。

9、數位發展部，公部門人工智慧應用參考手冊（草案），2025 年 1

月 8 日，網址：
https://moda.gov.tw/digital-affairs/digital-service/guide/15002，最後瀏覽日期：2025 年 3 月 18 日。

10、數位發展部資通安全署，公務機關全面禁用 DeepSeek AI 服務，未限制一般民間使用，數位發展部資通安全署新聞稿，2025 年 2 月 20 日，網址：
https://moda.gov.tw/ACS/press/news/press/15296，最後瀏覽日期：2025 年 3 月 21 日。

11、廖明輝，自由開講》主權 AI 崛起 TAIDE 捍衛數位主權，自由時報，2024 年 9 月 25 日，網址：
https://talk.ltn.com.tw/article/breakingnews/4809861，最後瀏覽日期：2025 年 3 月 28 日。

12、蘇思云，AI 基本法草案 數發部：在創新與風險中找平衡，中央通訊社，2025 年 4 月 23 日，網址：
https://www.cna.com.tw/news/afe/202504230387.aspx，最後瀏覽日期：2025 年 5 月 1 日。

13、AI 產品與系統評測中心，成立背景，網址：
https://www.aiec.org.tw/AboutUs/EstablishmentBackground，最後瀏覽日期：2025 年 3 月 18 日。

14、AI 產品與系統評測中心，關於 AI 驗證機構，網址：
https://www.aiec.org.tw/AICertificationBody/Index，最後瀏覽日期：2025 年 3 月 31 日。

15、2023 年廣島 G7 峰會領導人聯合聲明（G7 Hiroshima Leaders' Communiqué），日本外務省網頁，2023 年 5 月 20 日，網址：
https://www.mofa.go.jp/policy/economy/summit/hiroshima23/documents/pdf/Leaders_Communique_01_en.pdf，最後瀏覽日期：2025 年 3 月 20 日。其中涉及人工智慧部分，在該聲明第 38 段。

16、2023 年 G20 印度峰會領袖宣言（G20 New Delhi Leaders' Declaration)，印度政府外交部網頁，2023 年 9 月 9 日-10 日，網址：
https://www.mea.gov.in/Images/CPV/G20-New-Delhi-Leaders-Declaration.pdf，最後瀏覽日期：2025 年 3 月 20 日。

17、Bsi，ISO/IEC 42001 人工智慧管理－塑造值得信賴的人工智慧（AI）基礎，網址：
https://www.bsigroup.com/zh-TW/standards/isoiec-42001/，最後瀏覽日期：2025 年 3 月 14 日。

18、European Commission，Ethics guidelines for trustworthy AI，2019 年 4 月 8 日，網址：
https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai，最後瀏覽日期：2025 年 3 月 28 日。

19、European Parliament，Artificial intelligence act and regulatory sandboxes，網址：
https://www.europarl.europa.eu/RegData/etudes/BRIE/2022/733544/EPRS_BRI(2022)733544_EN.pdf，最後瀏覽日期：2025 年 3 月 24 日。

20、ELYSEE，Statement on Inclusive and Sustainable Artificial Intelligence for People and the Planet.，2025 年 2 月 11 日，網址：
https://www.elysee.fr/en/emmanuel-macron/2025/02/11/statement-on-inclusive-and-sustainable-artificial-intelligence-for-people-and-the-planet，最後瀏覽日期：2025 年 3 月 18 日。

21、Federal Trade Commission，The Federal Trade Commission 2023 Privacy and Data Security Update，2024 年 3 月 21 日，網址：
https://www.ftc.gov/system/files/ftc_gov/pdf/2024.03.21

-PrivacyandDataSecurityUpdate-508.pdf，最後瀏覽日期：2025年5月1日。

22、IMDA，Singapore’s first generative AI Sandbox to familiarise and help SMEs get head start in capturing new AI opportunities，網址，https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/press-releases/2024/sg-first-genai-sandbox-for-smes?utm_source=chatgpt.com，2024年2月7日，最後瀏覽日期：2025年3月20日。

23、IMDA，Singapore launches world’s first AI testing framework and toolkit to promote transparency; Invites companies to pilot and contribute to international standards development，2022年5月25日，網址：https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/press-releases/2022/sg-launches-worlds-first-ai-testing-framework-and-toolkit-to-promote-transparency，最後瀏覽日期：2025年3月25日。

24、IMDA，Singapore launches AI Verify Foundation to shape the future of international AI standards through collaboration，2023年6月7日，網址：https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/press-releases/2023/singapore-launches-ai-verify-foundation，最後瀏覽日期：2025年3月25日。

25、NIST，AI Risk Management Framework，網址：https://www.nist.gov/itl/ai-risk-management-framework，最後瀏覽日期：2025年3月21日。

26、OECD，AI principles，網址：https://www.oecd.org/en/topics/ai-principles.html，最後瀏覽日期：2025年3月5日。

27、PDPC，Singapore’s Approach to AI Governance，網址：https://www.pdpc.gov.sg/help-and-resources/2020/01/model-ai-governance-framework，最後瀏覽日期：2025 年 3 月 20 日。

28、UNESCO，UNESCO member states adopt the first ever global agreement on the Ethics of Artificial Intelligence，2021 年 11 月 25 日，網址：https://www.unesco.org/en/articles/unesco-member-states-adopt-first-ever-global-agreement-ethics-artificial-intelligence，最後瀏覽日期：2025 年 3 月 25 日。聯合國教育科學及文化組織網頁，網址：https://unesdoc.unesco.org/ark:/48223/pf0000373434，最後瀏覽日期：2025 年 3 月 25 日。

29、Whitehouse，Guidance for Regulation of Artificial Intelligence Applications，2020 年 11 月 17 日，網址：https://www.whitehouse.gov/wp-content/uploads/2020/11/M-21-06.pdf，最後瀏覽日期：2025 年 3 月 13 日。

30、Whitehouse，Removing Barriers to American Leadership in Artificial Intelligence，2025 年 1 月 23 日，網址：https://www.whitehouse.gov/presidential-actions/2025/01/removing-barriers-to-american-leadership-in-artificial-intelligence/，最後瀏覽日期：2025 年 3 月 5 日。

附錄一：「AI 人工智慧治理問題研析－以建構可信任 AI 為中心」專題研究報告（初稿）座談會紀錄及參採情形

時間：114 年 5 月 9 日（星期五）上午 10 時

地點：立法院法制局 304 會議室

主席：黃組長良瑩　　紀錄：李淑瓊

參加人員：

一、學者專家

東吳大學法學院　趙助理教授萃文

二、機關代表：

數位發展部

視察　蔡帛軒

專員　黃哲修

分析師　闕于閎

數位發展部數位產業署

科長　蔡欣瑩

數位發展部資通安全署

視察　劉禹君

國家資通安全研究院

經理　彭敏君

三、本局出席人員：

林研究員鈺琪

傅副研究員朝文

陳助理研究員樂庭

發言要點：

東吳大學法學院趙助理教授萃文

本座談會報告以「建構可信任 AI」為核心，從國家安全、語言文化、公共治理與國際法制等多面向研析 AI 治理的重要性與挑戰。AI 技術日益滲透社會各層面，尤其在軍事、國安、選舉與認知作戰中的應用，引發資安與主權風險。中國利用生成式 AI 模型強化數位主權，透過開放模型與話語權操作，推動極權價值的國際擴散，構成對民主制度與自由價值的潛在威脅。

AI 治理不能僅限個資與隱私風險，而應納入對民主制度、人類尊嚴與文化多元的保護。AI 若應用於軍武、自主武器系統，將可能導致人類失控的災難性後果，應透過國際合作建立相關規範。我國應參考歐盟人工智慧法與美國行政命令等相關監管機制，在鼓勵創新與保障人權之間取得平衡。並應積極發展「主權 AI」，建立以本土語料與語言文化為基礎的語言模型，避免外國模型影響我國政策與文化認同。AI 的發展應以人為本，強化透明度、可解釋性與問責機制，藉由正當法律程序確保公民權利與民主韌性不受侵蝕。唯有建構具責任、尊重人權且能應對未來挑戰的 AI 治理架構，方能實現 AI 為人類共同福祉服務的終極價值。

（趙助理教授另提供之書面意見請參見附錄二）

參採情形：

趙助理教授之意見係探討 AI 技術對國家安全與社會的影響，及 AI 濫用對全球構成安全與倫理之威脅，並主張發展主權 AI 以維護國家利益，強化對軍武應用、基礎設施及數位主權的法律監管，並呼籲應思考如何針對 AI 先進技術，基於國際規則，制定有效的全球規範，以維持秩序，其內容亦有贊同本報告論述者，爰錄供參考。

數位發展部

一、有關 AI 治理與監管架構，本部觀察與本案報告類似，目前國際間尚未有明確共識。有關報告建議我國優先採取美國較具彈性之

監管模式，與目前預計做法大致相符。本部將依據人工智慧基本法草案相關規定，推動「人工智慧風險分類框架」，描述 AI 應用上常見的風險，以供各目的事業主管機關，對其所涉領域之 AI 應用進行風險識別與評估，並視需要訂定相關管理規範。

二、為協助國內 AI 產業發展，「AI 產品與系統評測中心(AIEC)」目前以語言模型為評測對象，參酌國際 ISO 標準以及美國 NIST AI RMF 等相關規範，擬定評測項目包含公平性、準確性、可靠性、隱私及資安，並積極鼓勵業者送測；另 AI 測試實驗室及 AI 驗證機構分別以 115 年及 116 年通過 TAF 為目標。有關報告中建議「新加坡的 AI Verify 在推動國際合作與開源社群方面具有優勢，值得台灣借鏡。AIEC 可考慮引入開源工具與框架，提升測試的透明度與可驗證性，並加強與國際社群的連結，提升台灣在全球 AI 治理領域的影響力。」，AIEC 所使用測試框架即參考 NIST 開源 Dioptra 進行設計，並定期與 NIST 交流，亦於 2024 年與新加坡 AI Verify 成員對模型測試 Benchmark 互動討論，未來將持續參考國際相關標準與開源工具並與相關組織、社群交流，持續精進國內 AI 測試平台與方法。

三、現階段公務機關不得使用 Deepseek AI 等大陸廠牌資通訊產品，倘因業務需求且無其他替代方案，應具體敘明理由，經機關資安長及其上級機關資安長逐級核可，函報數位發展部核定後，以專案方式購置列冊管理，及遵守以下規定：指定特定區域及特定人員使用、使用理由消失應立即停止使用、以不含個資及資料的電腦單機將其下載並以斷網或非公務網路之獨立網路使用較為安全。

四、可信任的 AI，應該明確標的，關於可信任這個議題，認為應該分成兩個面向思考輸出內容正確性及 AI 模型直接造成的危害，在輸出內容的正確性部分，目前大部分針對模型的測試皆屬於這塊，然而內容輸出正確性的部分，以資安院測試過的 deepseek 為例，確實提供目的性偏頗以及提供不良（惡意、致危險）等輸

出，可能影響使用者之價值判斷。在 AI 模型直接造成危害部分，目前已知的 AI 模型，在尚未整合進其他系統或機制前，尚無能直接影響真實世界的 AI 模型。

五、規範應針對廠商要求提供技術細節，以防廠商透過自主研發惡意包裝有問題 AI 模型，造成損害。

六、為支援我國主權 AI 發展，數發部已研提「建構臺灣主權 AI 訓練語料庫行動計畫」，加速打造臺灣主權 AI 訓練語料庫，以系統性徵集具臺灣文化特色及語意連貫之高品質正體中文語料。

七、科技變革速度甚鉅，AI 技術實作機制宜採政策計畫逐年精進，扣合技術發展，以法律明文規定「主權人工智慧訓練資料庫」之技術平臺文字，恐有失彈性，原草案第 15 條應已達政府須致力主權 AI 發展之效果，爰 AI 基本法草案第 15 條所建議文字建請移除。

參採情形：

一、第一點至第六點意見，未反對本報告論述，錄供參考。

二、第七點意見有關「……以法律明文規定『主權人工智慧訓練資料庫』之技術平臺文字，恐有失彈性……」一節，已參採相關意見，將原建議於人工智慧基本法草案第 15 條增訂 1 項，改為酌修第 2 項文字為「政府應致力提升我國人工智慧使用資料之品質與數量，收集本國語言、文化、法律等相關資料，提供人工智慧模型訓練運用，確保訓練結果維繫國家多元文化價值與維護智慧財產權。」(頁 31)。

林研究員鈺琪

一、本報告提到短期內應採取較具彈性的監管模式，以促進創新並降低企業的法律遵循負擔(頁 22-23)，然而，彈性監管如何具體實施？例如在某些高度風險領域，單純的彈性監管可能無法充分保護公共安全與個人隱私。爰建議或可進一步設立區分性監管機制，對不同領域、不同風險等級的 AI 技術設定不同的監管要求，

對於具有高度社會影響的 AI 應用（如金融、醫療等），應加強監管與合規要求，對於風險較低的 AI 應用，則可採取較為寬鬆的監管政策。

二、在「發展主權 AI 以支持可信任的 AI」部分，本報告提出推動「主權 AI」的發展，尤其是發展本土語言模型（如 TAIDE）以強化數位主權與文化保護。然而，除了語言模型的開發，或許還需要在資料治理、算法透明、AI 倫理等方面做長期規劃，例如，如何利用本地語言資料進行大規模語言模型訓練，如何處理來自不同領域（如金融、醫療）的敏感數據以保障個人隱私等。因此，除本報告建議將主權 AI 法制化外，在技術標準的制定、跨部門協調、國際合作等方面，亦是未來政府應進行安排與規劃之處。

三、新加坡於 2023 年發表「國家 AI 策略 2.0」(NAIS 2.0)，提出「AI for the Public Good」為願景，政府投入約 5,200 萬美元啟動「SEA-LION」計畫，開發涵蓋東南亞語言的大型語言模型，並建立 SingaBERT 支援多語種本地語言處理，透過 AI@SG 計畫整合超級電腦與應用測試平台，推動 AI 研究。有鑒於新加坡致力於發展主權 AI，以強化國家數位能力與全球治理影響力，本報告亦強調我國發展主權 AI 之必要性，並提出法制化之建議，爰建議或可適度簡介新加坡發展主權 AI 之相關策略，以作為我國之參考借鏡。

參採情形：

一、第一點意見，有關「……彈性監管如何具體實施？……爰建議或可進一步設立區分性監管機制，對不同領域、不同風險等級的 AI 技術設定不同的監管要求……」一節，按我國有關人工智慧之立法進程，尚處於主管機關草擬基本法法案階段，前述建議似較宜規範於各目的事業主管機關對其主管業務所制定之作用法中，爰錄供參考。

二、第二點意見係提及政府未來就 AI 監管應進行之安排與規劃之處，未反對本報告論述，錄供參考。

三、第三點意見，按我國人口約有 2,300 餘萬人，而新加坡人口僅有約 590 萬人，故有關主權 AI 之發展，我國不論是資料規模或國情條件皆與新加坡差異甚鉅，政策取向亦會不同，是以，新加坡發展主權 AI 之相關策略，無法直接援引作為我國之借鏡，爰錄供參考。

傅副研究員朝文

一、本報告係針對「AI 人工智慧治理問題研析－以建構可信任 AI 為中心」進行探討，其體系完整、論述詳實，所建議方向皆深具參考價值。

二、本報告問題研析與建議之第一點有關 AI 治理之歐盟與美國監管模式部分，建議在立法上應優先採取美國式發展導向、彈性自律的監管模式，以鼓勵企業自律與創新。惟歐盟對 AI 治理所以採取較為嚴格之監管模式，係因其欠缺如美國 CHAPGPT 等 AI 產業。蓋在未來 AI 時代，數據資料就是財富，不宜任由外資企業無償自由運用，甚至反向塑造有利於外資的 AI 應用環境。我國有關 AI 產業主要在於伺服器等硬體製造部份，軟體服務部分並非我國企業專長，反而係屬外資企業壟斷範圍。簡言之，我國所面對之 AI 環境與歐盟較為類似，為維護我國數據資料之運用安全，爰建議在政策制定上應先確認採取歐盟監管模式，做為訂定後續 AI 治理法規的基礎。

參採情形：

一、第一點意見，未反對本報告論述，錄供參考。

二、第二點意見，按依歐盟之監管模式，企業需投入高額法律遵行與風險評估之成本，對中小企業與新創公司而言，將造成其沉重負擔，又我國 AI 產業仍處於起步階段，且企業規模以中小企業為主，市場規模不若歐盟龐大，若實施嚴格監管，恐阻礙產業創新，不利 AI 產業發展，爰錄供參考。

陳助理研究員樂庭

一、本報告「貳、人工智慧的發展背景與可信任 AI 的概念」，有關國科會預告制定人工智慧基本法草案一節，行政院本年 2 月 26 日已指示 AI 基本法草案改由數發部主政推動法案立法及續行條文解釋 ，建議本報告可補充相關資訊。

二、本報告問題研析與建議「一、我國推動可信任 AI 立法策略之比較與評估」中建議我國主管機關應先確定監管路線後再進行法規草擬一節，考量 AI 所致風險具有不同之侵害程度，相關影響之歸責應以風險大小分級，研議不同規範密度已建構可信賴的 AI 應用環境，數位發展部刻依據人工智慧基本法草案訂定人工智慧風險分類框架草案，以供各目的事業主管機關對其所涉領域之 AI 應用進行風險識別與評估，並視 AI 應用風險之需要訂定管理規範，謹提供本資訊供撰稿人參考。

三、本報告「參、外國立法例及國際規範比較分析」，謹補充歐盟規範動態供參：歐盟執委會業於 2025 年 2 月 4 日發布歐盟執委會關於 2024/1689 號規範所設立的禁止人工智慧行為指引，就 AIA 第 5 條「禁止的 AI 行為」之各禁止事項分別闡述其立法理由、禁止行為之具體內涵、構成要件、豁免適用之特定情形等，並以例示說明，以提高 AIA 的法律明確性，並確保歐盟區主管機關執法的一致性 。

參採情形：

一、第一點意見，本報告已於註 1 敘明「……行政院 2025 年 2 月已指示 AI 基本法草案改由數位發展部主政推動法案立法及續行條文解釋……」。

二、第二點及第三點意見係敘及數位發展部刻依據人工智慧基本法草案訂定人工智慧風險分類框架草案，及歐盟執委會於 2025 年 2 月 4 日發布歐盟執委會關於 2024/1689 號規範所設立的禁止人工智慧行為指引，未反對本報告論述，錄供參考。

散會：上午 10 時 40 分

附錄二：東吳大學法學院趙助理教授萃文提供之書面意見

壹、前言

AI 被應用在越來越多的領域，可謂長驅直入社會各個角落，說現今社會為 Al 社會，也不為過；惟 AI 技術亦可應用在戰爭武器、國安情資、選舉，乃至應用於認知作戰，擾亂資訊判讀，影響人們的思維，威權政體更用來鞏固政權。因此，對國家安全而言，在支持科技創新與發展的同時，能夠保護國家和人民免受科技可能造成的潛在危害，AI 平臺需要健康有序之政策與法律；該如何建立監理防護機制，對其為適切法律規制，以確保能安全確當應用，係非常困難之議題，但實為重中之重 。大院法制局李研究員淑瓊完成專題報告，個人有幸參與座談見學，就報告提供一些不成熟建議，就教於各位方家與先進。

貳、問題研析

一、中國大陸透過 AI 服務影響全世界的價值觀

中國對於生成 AI 大型語言模型，借力使力採取最省錢的捷徑用既有最新模型，其主要目標在於持續運用開放模型權重，讓新進者在其語言模型基礎上精進研發，使其 AI 模型能不斷進化。更重要的是，中國對於「數位主權」的主張，勢將滲透擴散，藉由全球對其關注熱度不斷增升，而迅速擴大使用數量，讓輸入資料與訓練成熟度持續獲得挹注，確立中國在全球市場立足；在習近平「人類命運共同體」與構建「網路空間命運共同體」的大旗之下，持續對全球 AI 治理與監管領域的標準及指導綱領下手，以共同發展與共享成果名義，讓其社會主義模式極權價值觀，散發的影響日益強化與深化，並以此抗衡以美國為首的民主價值一致的 AI 通用或專用多模式模型，進而運用推理模式生成符合其意識形態與極權價值觀的文化與歷史詮釋，形同將大型 AI 語言模型轉化為干預影響利器，以反制、詆毀並進一步扭曲操弄自由民主價值。

中國在提倡 AI 治理上，不斷運用與歐盟相近的論述話語，強調全球 AI 發展，應依循可解釋、可信賴原則，讓發展成果開源且可共享，

攻占全球AI治理的話語權。面對中共全力支持發展AI，美國勸說歐洲盟友勿以過度嚴苛的個資法與AI風險治理，大幅侷限AI的研究發展與應用推廣，避免讓中共累積後發先至的態勢。

美國根據中國制定《反間諜法》與《國家情報法》等規範，明定中國企業有配合蒐集國家安全相關情報之義務，研判中企無法拒絕政府蒐集情資之要求，恐產生無法預測之風險，因而祭出反制行動。觀諸人權與倫理議題一直是民主國家對華為與中興通訊進行制裁之緣由，可為例證。

二、AI遭濫用須從全球角度思考因應戰略

聯合國秘書長古特瑞斯（Antonio Guterres）指出：因為AI被濫用在網路攻擊、深偽、散播虛假訊息與仇恨言論，恐對全球和平與各國安全造成重大影響；除AI系統存有潛在不穩定或故障外，AI結合核子、生化武器或機器人更是令人擔憂，恐對各國安全與世界和平造成重大危害。面對當前變局，須從全球角度去重新思考因應戰略，故聯合國有必要推動建立新國際準則、簽訂新國際條約，建置相應的全球防護機構。

三、嚴肅看待AI造成的生存威脅之可能性

AI應用於軍武的發展對人類生存的威脅，目前還只是推測性的擔憂嗎？在不對稱的 AI 戰爭能力中，技術先進的國家可能通過使用AI在軍武發展與認知作戰中獲得顯著優勢，從而加劇全球權力不平衡；這種不對稱性，可能讓較弱小國家更難以保護自己免受AI影響，聯合國「人工智慧倫理建議書」亦指出人工智慧正在加劇原已分裂和不平等的世界，應該加以控制。面對一個AI越成功，也是人類越來越陷入險境的驚悚未來，科技巨擘馬斯克對於 Open AI 的發展就很焦慮，警告人工智能越成功，人類就越危險，到最後就會瀕臨失控，而成功的科學突破，卻成為人文世界的災難。

孔祥重院士指出，人類和AI都很危險，人類的危險是殺人、放炸彈這種明顯可見的危險；但AI的危險並不直接可見，他可能把整個社會的文化價值都改變卻不被察覺，這是更可怕的災難。我們必須確保AI數據與人類從歷史中所獲得的智慧與教訓保持一致，而這種

一致性可以從人文主義的教育中去實踐；人文領域的人才，可以「把AI 變得更安全」，可以「確保 AI 為我們做正確的事」 。

四、促進 AI 公共化創造人類共同善

隨著愈來愈多部門將 AI 用於與國家安全有關領域，如利用 AI 監控關鍵基礎建設等；為確保 AI 技術不會傷害人類，宜加強推動系統整合，確保能緊急啟動備援通訊系統，並納入高度數位韌性的設計規範，以期各項目在實際應用時具備最佳穩定性，讓政府單位能正常運作，避免依賴以 AI 進行決策之結果出現偏差，維持社會基本生活 ，達致「友好人工智慧」系統安全所追求目標。此實有賴以正當程序保障公民權利，提升透明度要求，測試評估的展開，說明理由之引入等，藉由正當程序原則之約制，避免公民權利遭受威脅 。

五、AI 發展應以人為本創造人類共同善

AI 長久發展下去，恐怕世界人類將會同歸於盡，為應對 AI 的風險與挑戰，AI 的研發與應用必須要監管，充實其安全措施，帶給人們的是進化的人類，成就善的共業，成就完美的生命共相，讓世界永續和平，人類永續的幸福；這是人類生命的目的，也是 AI 的終極價值。建構安全且有效使用 AI 的社會，唯有「以人為本」的 AI 社會原則，才能達到公共化，實現 AI 科技最終所要的價值目標。學者林昕璇以國家責任角度，主張各國對於 AI 自主性武器之研發與擴散確屬不可逆，應提高國家責任之歸責門檻為限制無過失責任(limited strict liability)，希望緩解軍武科技外溢後產生的監管難題。

參、AI 法規治理規範模式建議

一、資安即國安

AI、量子運算已無關地緣，基於國安考量，科技發展不得不管制。報告提出，基於國家戰略需求，應發展「主權 AI」，由國家主導或支持自主研發、建置及運用之人工智慧系統，包括資料收集、模型訓練及技術應用等，確保國家安全、經濟發展和文化自主性，以支持可信任的 AI。科技戰不可免，台版 Chat GPT 何在？發展以我國本土語料與語言習慣訓練之大型語言模型，不僅能強化政策溝通與在地服務，也能作為語言文化保存工具；AI 開發大抵要以億萬美元起跳，大

型語言模型要投入數十億美元，才可能完成。我國已錯失開發大型語文模型的最佳良機，要用美國的，問題我們能管制美國嗎？自主研發年輕人會用嗎？

從主權 AI 的層面來看，擁有我們自己的正體中文大型語言模型，對未來的技術發展非常重要，現下政府願投下成本，已與輝達等頂尖 AI 晶片巨擘合作，打造 AI 資料中心及超級電腦，是種大型生成式 AI 語言模型，用於訓練我國的「可信任 AI 對話引擎」，為時稍晚，但殊值讚賞！整體而言， 科技牆應該多寬、多高、多厚？有待隨時檢討，滾動修調。

（一）網駭竊密

溯源駭侵源頭、方式，發覺漏洞，建立防火牆。

（二）爭議訊息

即時阻絕爭議訊息再擴散，但過多爭議訊息資料，要如何防制？防杜假新聞並無太大效益，AI 不是人，處罰 AI 幹嘛！只能透過自律，誰有能力管網路。兒少和色情問題各國大多都有規管，政治性議題較不會納管，要求封鎖境外網站，增加過濾機制，就得補助業者經費。真要納管，恐怕只得學中國大陸，問題在法治先進國家大抵認中國欠缺法治精神；更且，中國若真監管得住，又怎麼有那麼多翻牆的？

（三）關鍵基礎設施

從關鍵基礎設施面向出發，機先建立防護網阻絕駭侵漏洞，如能源、水電、交通與金融系統等類 AI 模組，潛藏諸多風險，為確保系統的可控性與應變能力，或可透過立法要求關鍵基礎設施 AI 採用主權模型或具備透明性、安全性之系統，並建立跨部會監管架構。

二、國家安全與網路隱私

從數位主權概念言，「主權者」為國家公權力或私人主體？數位主權其意涵的探討及規範的連結，厥為個人資料保護的基本權利、資訊隱私權或資訊自決權。

科技巨擘對網路世界與數位平台使用者的全面性資料蒐集分析

與處理，已非僅止於商業目的（廣告投放），更用於干預並影響人的各種行為舉止（政治意向及投票行為），數位主權即在彰顯，人們越來越透明，行為舉止越來越可受操控。

各國目前對於 AI 治理與法制的討論，大抵偏向 AI 衍生問題的課責機制，對於事前管制較少著墨。即便是事前管制，亦多聚焦在資料及隱私治理的面向，把個資保護與資通安全作為管控 AI 風險的主要考量。針對創成危險的特殊應用，如在軍武系統、關鍵基礎科技與國家安全等特定 AI 發展與應用上，反而有立法管制的急迫性。

三、AI 規管模式

我國 AI 立法應綜合考量產業發展與政策目標，選擇最適合監管模式 。在借助 AI 系統同時，AI 應用存在的安全、歧視和隱私風險，應納入法治軌道。歐盟執委會人工智慧高級專家小組 2019 年提出《可信任 AI 倫理準則》，落實在 AI 整個生命週期的持續稽核，藉由遵守人性、法治國及不歧視等各項倫理原則，並透過「以人為本」的作法，達致讓 AI 在合法之外，猶能符合倫理及穩健的高品質要求，該準則具體描述可信賴的 AI 包含三大要求七大面向，國家科學及技術委員會預告制定「人工智慧基本法」草案，是否符合上開要求。

歐盟於 2024 年 3 月 13 日通過全球首部《人工智慧法案》(Artificial Intelligence Act)，希望降低因 AI 技術快速發展產生之族群偏見、侵犯隱私與其他各種風險，避免不適當或惡意的設計、發展、部署和使用所帶來的風險及損害，削弱對人權、自由與生命的保護，共同提升 AI 系統的安全及可信度 ；將潛意識操縱、社會信用評分、大規模遠程人臉辨別列為 AI 嚴禁發展事項，從功能上為生成式 AI 設限，即便失控亦可將損害降至最低，甚而消弭於無形 。

歐盟對於人工智慧規範架構之推動與進程相對完整，其 AI 法草案雖在幾個關鍵議題上爭論仍多，如關於不可接受風險、高風險 AI 規定詳盡，但卻明文排除可能更嚴重侵害歐盟基本價值觀及人民基本權利之軍事用 AI。有別於歐盟，美國 AI 總統行政命令，將 AI 風險聚焦在國家安全、經濟與公民權利等特定風險議題，要求聯邦機關為 AI 技術制定標準與提供使用指引，並賦予企業應向聯邦政府通報資

訊之義務，政策發展轉向強調倫理與負責任的AI，以更積極的態度參與AI治理，要求美國公私部門以安全、可信賴的方式發展及使用AI，認為AI應遵守倫理標準並強化社會責任 。

本報告指出，鑒於我國AI產業主要市場集中於少數技術領先國家，且出口面臨地緣政治與國安挑戰，應優先採取美國式發展導向、彈性自律的監管模式，以鼓勵企業自律與創新，並保留政策調整空間，殊值贊同。

面對AI之發展與應用可能造成的風險，AI的技術特性或利用方式，不僅只會對個人資料產生危害，更可能會違反正當法律程序、加深偏見或侵害其他基本權利，是以，要確保AI對人類不會造成負面衝擊，除個資危害風險以外，還得多關注其他類型的風險。AI法制與個資法制互為關連，惟聚焦個人資料保護並無法取代AI法制，二者或可連動規制。

四、建構我國AI治理準則

為降低AI科技發展帶來之潛在負面影響，確保AI發展與安全，將AI系統的安全性視為資訊安全的重要一環，政府之監管思維必須由初期的自律為主，轉為朝向負責、公平、可靠與可治理的他律監管。透過強制性之法律規範，結合事前預防、事後究責措施，確保AI技術在負責任的監管成長，避免科技發展失控，同時有賴網路安全、數據安全及智慧財產權等其他新興技術穩定引導下，讓人工智慧在平等、安全及隱私下應用，看見一個更安全、公正及創新的AI未來，以確保人類生存及生活的條件 。

截至目前為止，各國對AI監管的法律框架仍有許多缺陷，現有規範大多聚焦在使用倫理、隱私保護與決策透明度，缺乏專門針對AI在軍武研究的風險管控。聯合國教科文組織倡議通過的一項全球協議「人工智慧倫理建議書」，對照數位發展部偕同工業技術研究院與國家資通安全研究院等機關展開AI評測工作，逐步建立評測制度與規範，2023年12月成立AI產品與系統評測中心，建置相關制度、標準及評測體系；建立包含安全性、可解釋性、彈性、公平性、準確性、透明性、當責性、可靠性、資料隱私、系統安全等十項評測標

準，將風險分成對人、對企業、對生態系統的三大潛在危害 ，在使用和開發過程中遵循對人權、隱私之保護，除確維產品和系統的性能，並確保其符合道德和法規要求，提高民眾對 AI 的信任。

至於主責機關權責分工可以規劃設計得更細緻一些，國科會對科技人才及預算資源的掌握，或許較能發揮統合作用和力道。像國家安全法由國防部、內政部和法務部三部共同為業管機關，但國安情報的統合指導及協作卻是國家安全局，有些名實不符。另為統合政府各部門資源及意見，仿歐盟治理架構成立「人工智慧委員會」作為協調平台暨諮詢機構。

五、讓人立於更有價值的位置

現下 AI 系統從如何取得及輸入個案數據、風險群組之建立與正確性的評估等事項，均取決於人的價值判斷及主觀決定；主流 AI 技術所運用之原理與大數據及資料分析有關，在創作歷程中所扮演之角色，大抵為人類之輔助工具，真正主導者是人。因此，如何設計新型態之監管和治理措施，關鍵即在數位信任的建構與維持，提升全球的發展環境，終極實現全球智慧社會的宏偉願景。

鑑諸 AI 及其衍生運用，當前一般認為仍係作為一個在定義內涵與界線範疇上，仍持續不斷變動而未臻明確的技術集合體，解釋論或立法論上，均仍存在高度不確定，連帶致令國際上未能形成共識機制，屬於一項極具新穎性和與未知程度甚高的規範秩序與規則體系，國際社會亦需加強合作，制定適應性強且明確的規範標準，惟 AI 科技並無「道德者」或「愛國者」的位置，可能會做出與人性相悖的演算，這是關乎人類生存策略的問題。

AI 系統的創立宗旨在「為全人類服務」，其發展要非是一個斷面，會不斷創新，要與時俱進，隨應變局，積極推動安全發展指引外，建構全球性開發治理機制，規範政府活動，限制政府權力，建立有責的政府和有義務的業者，構築一個「人類的公產」之開發和創新規制，用「法」來保全人類，實現文明社會的最大公約數。任一 AI 技術的發展，都有隱私、資料安全及倫理問題需要克服，重點在科技與人文交織的過程中，人類確實需要謹慎前行，以確保 AI 的進步能夠真

正服務於人，而非帶來新的矛盾與傷害。

肆、現行 AI 治理之國際規範及多邊倡議大抵針對一般應用

軍武應用的治理與一般應用的治理，事涉不同法制框架的複雜度，監督機制、具體作法，乃至如何妥適平衡與良善治理方式，以及足夠的透明度規範，無不需要各方對話，期能達致 AI 軍武應用治理的一致性與有效性，以促近法遵意願。

現行 AI 治理之國際規範及多邊倡議，大多屬指導方針，提供一個健全治理政策和程序；其核心原則包括：尊重人權與人類尊嚴、維護環境與永續發展、促進和平與社會正義、公平取用技術及性別平等與文化多元性。這確實有助於最大限度地減少與人工智慧相關的潛在風險和負面影響，並提出建立全球 AI 治理架構之必要性；惟此大抵針對一般應用，針對 AI 新武器崛起涉及之國際規範秩序，非國際法社群在制定條約之際所能預見，針對 AI 新武器崛起，涉及之國際規範秩序，非國際法社群在制定條約之際所能預見。就軍武、戰爭等創成危險的特殊應用，治理及處置前置手段，才具有正當性與比例性，確保負責任使用，風險有效管控；宜立法禁止對人類社會有不可接受的風險之 AI 系統或應用，作為國際間強化 AI 治理的基礎。

伍、絕不容許讓 AI 變成滔天巨惡

人工智慧技術於國安及軍事領域快速發展，將持續推進戰爭型態與國防面貌，甚至從根本上改變「戰爭的本質」；AI 應用到戰爭、金融、醫療等產業時，需符合各該領域法規的制約外，有必要審視其對人類和平與安全的影響；尤其目前被用於戰爭的自主性武器，更應進一步評估影響及管制，一旦完全交由 AI 決定，不排除可能產生人類無法承受的錯誤；AI 軍武系統開發應建立可受控制的環境，保留人類主動權，確保 AI 符合人類要求與預期 。

綜言之，文明社會存在著永恆正確，不容侵犯的道德原則；這是人類文明進化的標誌。毀滅性「自主武器系統」，要非解決衝突的選項，更是嚴重悖德行為，是任何文明民族均無可規避的法則。由於自主武器系統擁有的軍備優勢和智慧化威力，任何具備開發能力的國家都不願在這個新興領域缺席，允宜建立 AI 軍武應用的國際共識與治

理。為了維持人類文明的前進，應制定禁止或限制使用《具有致命性之自主武器系統》的國際公約，嚴禁類此技術的發展行為。因此，必須轉而思考如何針對 AI 先進技術，基於國際規則，制定有效的全球規範，以維持秩序。

以上個人見解尚屬粗淺，有欠成熟，或未詳盡析論，謹為拋磚引玉，就教方家，以為後續深入思考與檢討的憑藉。