人工智慧風險管理的國際發展與應用

AI國際法遵需求迫在眉睫 人工智慧風險管理成顯學

AI 發展日新月異，憂心類似電影情節威脅到人類，歐盟與美國陸續公布在 AI 風險管理的內容，值得相關發展此應用的廠商及企業借鏡。

文／梁日誠

隨著**歐盟人工智慧法案（EU AI Act）**於歐洲議會獲得委員會層級的通過後，美國則在拜登總統令（EO 14110）的引領下，近日於商務部下成立了**U.S. AI Safety Institute Consortium（AISIC）**，匯整資源以建立安全與可信任的 AI，在可預見的將來，AI 科技的創新與衝擊將取得新的平衡，以使得消費者在享受 AI 先進產品與服務的同時，基本人權得以維護。

其中，在歐美 AI 相關法規的推動內容中，可以發現 AI 的風險與機會是衡量輕重（例如，不同高低風險等級的適法性）的共同點，而 AI 風險管理則顯現於相關法規內容中，也將成為合規的具體要求。

歐盟 AI 相關法規推動狀況

從歐盟的 AI 法規化推動進程方面來看，EU AI Act 於 Article 9 Risk Management System 與 Article 17 Quality Management System Para.1.(g) 對風險管理系統提出要求，歐盟執委會進一步以 2023-05-22 的執行決定要求 CEN 與 CENELEC 偕同 ISO（依據維也納協議）、IEC（依據法蘭克福協議）及 ETSI 制定 AI 相關歐盟標準（共 10 項）以支持歐盟的 AI 政策。

在 **CEN／CENELEC／JTC 21** 的工作計畫所顯示的最新進度中，相關於 Article 9 的 EN ISO／IEC 23894：2024 Information Technology - Artificial Intelligence - Guidance on Risk Management 已經核准，為等同採用 ISO&IEC標準；至於相關於 Article 17 的 prEN ISO／IEC 42001 Information Technology - Artificial Intelligence - Management System 則處於發展階段，此亦為等同採用 ISO＆IEC 標準，各涉及 EU AI Act 的組織可以做為合規機制的參考。

美國 AI 相關法規推動狀況

至於美國的 AI 法規化推動進程方面，在於 EO 14110 總統令中，要求依據 **AI Risk Management** Framework（NIST AI 100-1 AI RMF）進一步發展相關的資源及實作，並研議以 Secure Software Development Framework（NIST SP800-218 SSDF）來支持 Generative AI 與 Dual Use Foundation Models 的安全發展。

此外，NIST 初擬 NIST AI RMF 的可信任特性與 OECD Recommendation on AI、EU AI Act 間，及 **AI RMF** 與 **ISO 23894** 間的多個交互對應文件，供各界參考應用，若涉及美國市場的 AI 相關組織或可酌參，以規劃對應的合規機制。再者，相關於 **OECD Recommendation on AI** 的 OECD AI 風險管理互通性框架（AI RMIF）可見於 OECD Governing And Managing Risks Throughout The Lifecycle For Trustworthy AI 的文件。

NIST AI RMF	OECD AI RMIF	ISO 23894 （爰引 ISO 24028）
◇ Valid and Reliable ◇ Safe ◇ Secure and Resilient ◇ Accountable and Transparent ◇ Explainable and Interpretable ◇ Privacy-Enhanced ◇ Fair with Harmful Bias Managed	◇ Inclusive growth, sustainable development and well-being ◇ Human-centred values and fairness ◇ Transparency and explainability ◇ Robustness, security and safety ◇ Accountability	◇ Reliability ◇ Availability ◇ Resilience ◇ Security ◇ Privacy ◇ Safety ◇ Accountability ◇ Transparency ◇ Integrity ◇ Authenticity ◇ Quality ◇ Usability

<表一> AI 原則（Principles）。

NIST AI RMF	OECD AI RMIF	ISO 23894 （延伸於 ISO 31000）
◇ Govern ◇ Map ◇ Measure ◇ Manage	◇ Define ◇ Assess ◇ Treat ◇ Model	◇ Communication and consultation ◇ Scope, context and criteria ◇ Risk assessment ◇ Risk treatment ◇ Monitoring and review ◇ Recording and reporting

<表二> AIRMs功能

值得注意的是，**美國聯邦貿易委員會（FTC）**於 2023 年 11 月間，授權了 AI 相關產品與服務的強制程序條款，並於 2024 年 1 月公開提醒 AI 公司遵從其隱私與機密性承諾。

AI 風險管理（AIRM）漸受重視

目前來看，幾個主要經濟體的 AI 發展與法遵要求正同時並進，**AI 風險管理（AIRM）**也於其中扮演基礎的支撐角色，如何選擇適合的 AIRM 機制來支持 AI 業務並達到 AI 法規遵循，便是各組織的重要課題。目前 AI 領域經常被討論或採用的 AIRM 機制包含如 **NIST AI RMF、OECD AI RMIF、ISO 23894** 等，因應 AI 領域的風險管理特性，此些 AIRM（以下簡稱 AIRMs）機制也於以下幾個例舉的 AI 議題（如：AI 原則、AIRM 功能、AI 系統生命週期、AI 控制措施、AI 關注方）中異曲同工的展現支撐 AI 特性的共通性。分述如下：

AI 原則（Principles）

Trustworthy AI 是國際間所關注的 AI 須具備的特性，影響無法達到 Trustworthiness 的 AI 風險，便成為風險管理的焦點。AIRMs 將 Trustworthiness 相關的 AI 原則詮釋於其各自機制中，作為風險管理的目標或來源，**如<表一>**。

AIRM 功能（Functions）

AIRMs 的功能（Functions）或稱步驟（Steps）或過程（Processes）具有迭代（Iterative）的特性，持續地動態運轉與優化於組織中，並與組織內的其他的現有風險管理機制共存，綜整**如<表二>**。

NIST AI RMF	OECD AI RMIF	ISO 23894 （爰引 ISO 22989）
◇ Plan and design ◇ Collect and process data ◇ Build and use model ◇ Verify and Validate ◇ Deploy and use ◇ Operate and monitor ◇ Use or impacted by	◇ Design, data and models ◇ Verification and validation ◇ Deployment ◇ Operation and monitoring ◇ Retirement	◇ Inception ◇ Design and development ◇ Verification and Validation ◇ Deployment ◇ Operation and monitoring ◇ Continuous validation ◇ Re-evaluation ◇ Retirement

<表三> AI系統生命週期。

‖ AI 系統生命週期（Lifecycle）

AI 的各式產品、服務、機制與應用建立於 AI 系統之上，AI 系統具備生命週期的迭代特性並與資料生命週期（Data Lifecycle，參考 ISO 8183）與資料品質管理生命週期（Data Quality Management Lifecycle，參考 ISO 5259-3-發展中）交相並存，AI 各項議題須完整地涵蓋 AI 系統生命週期，AIRMs 的 AI 系統生命週期展現於**<表三>**。

‖ AI 控制措施（Controls）

選擇控制措施處理風險，是風險處理的選項之一，並須衡量風險與機會間的得失，控制措施的選擇與施作須完整地涵蓋 AI 系統生命週期與 AI 原則／Trustworthiness，方能有效的管理 AI 風險，AIRMs 相關的可考量 AI 控制措施說明於**<表四>**。

‖ AI 關注方（Interested Parties）

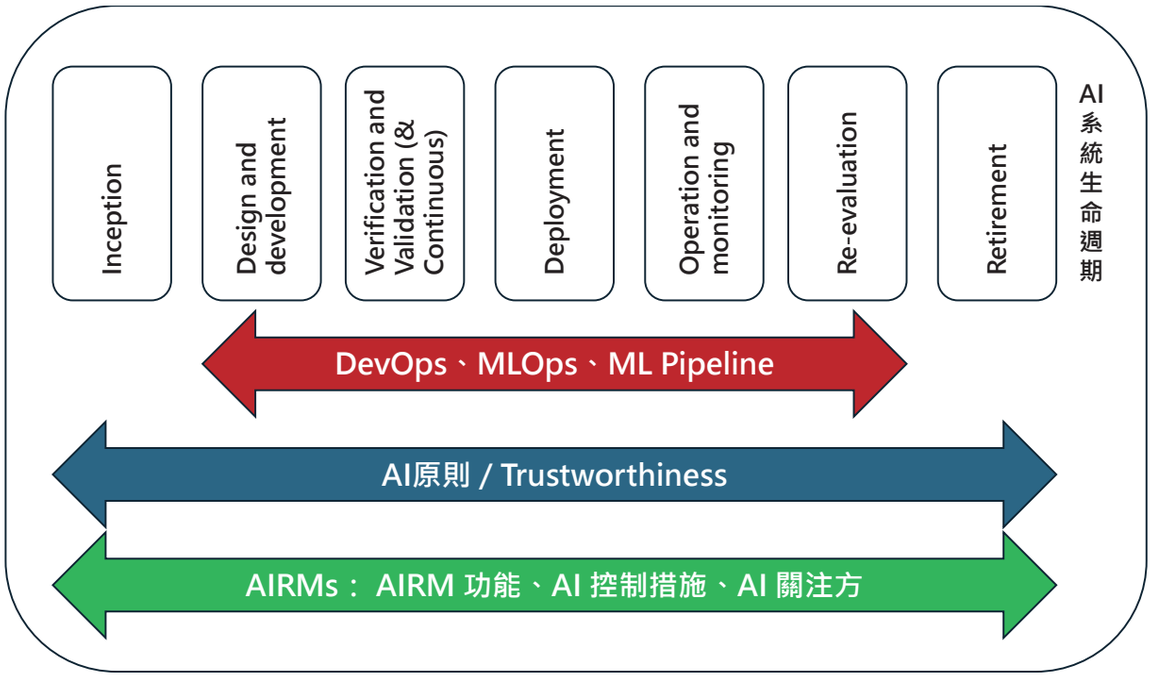
關注方與其對應的風險，須於 AI 系統生命週期中識別，包含如 Human-in／on-the-loop、Multi-Stakeholder Feedback 與供應鏈的情境，以兼顧關注方的外部威脅與內部威脅，AIRMs 的 AI 關注方列舉如下：

- NIST AI RMF：NIST AI RMF Appendix A 的各 AI Actor
- OECD AI RMIF：OECD Advancing Accountability in AI Chapter 2.3 Actors
- ISO 23894（爰引ISO 22989）：ISO 22989 Clause 5.19 AI Stakeholder Roles

NIST AI RMF	NIST 各現有或發展中的標準，如： ◇ NIST SP1270 Towards a Standard for Identifying and Managing Bias in Artificial Intelligence ◇ NIST SP800-218 SSDF ◇ NISTIR 8269 (draft) A Taxonomy and Terminology of Adversarial Machine Learning
OECD AI RMIF	OECD Advancing Accountability in AI 的 生命週期 Treat 階段中，對應於各 AI 原則與完整生命週期的 Process-related 或 Technical 方法（Approaches）
ISO 23894 （依 ISO 42001 選用）	ISO&IEC 各現有或發展中的標準，如： ◇ ISO 42001 Annex A/B 控制措施 ◇ ISO 24028 Overview of Trustworthiness in AI, Clause 10 Mitigation Measures ◇ ISO 5469 Artificial Intelligence - Functional Safety and AI Systems, Clause 10 Control and Mitigation Measures ◇ ISO 27090 Cybersecurity - AI（發展中） ◇ ISO 27091 AI Privacy Protection（發展中）

<表四>AI 控制措施。

以上例舉的各 AI 議題間的關聯性（AI 系統生命週期以 ISO 22989 為例），如**<圖一>**所示，若增／併考量其他的 AI 議題時，也須充分考量各 AI 議題間的完整關聯性。



<圖一> AI 議題間的關聯性

AIRM 在衝擊評鑑上的要求

進行 AI 風險管理時，經常會同時考量如 **AI 系統衝擊評鑑**（AI System Impact Assessment，參考 ISO 42005-發展中，為 ISO 42001 要求事項）、**人權衝擊評鑑**（Human Rights Impact Assessment）、**道德衝擊評鑑**（Ethics Impact Assessment）、**演算衝擊評鑑**（Algorithmic Impact Assessment）等的進行，以符合不同地區對 AI 治理的合規要求，此些評鑑也均具有迭代的特性，需要持續地動態管理與優化。

對於 AI 風險管理的威脅與脆弱點識別時（包含 Trustworthiness 的各特性），業界已經持續累積 AI 領域經驗可供參考，如：OWASP LLM Top 10（v1.1）、OWASP ML Top 10（v0.3）、OWASP Top 10 API Security Risks – 2023（若 AI 產品與服務涉及API）、MITRE ATLAS Machine Learning Threat Matrix 與由國際間多國合作產出的 Guidelines For Secure AI System Development 等，可供選擇控制措施時之借鏡。

在面對不同的市場或經濟體時，或有不同 AIRM 機制的選擇考量，以利合規的達成，惟考量 AIRMs 的共通性與互通性時，ISO&IEC 領域（AI 由 JTC1／SC42 主責）的 **AIRM ISO 23894** 搭配 **AIMS ISO 42001**，不失為一進可攻退可守的選擇，再輔以已公告與開發中的 AI 相關國際標準，包含 ISO&IEC 與歐盟 CEN／CENELEC／JTC 21 合作的國際／歐盟等同標準，資源及永續性最為豐沛。

另一方面，AIRM 與 AIMS 等國際標準可提供各國政府在制定 AI 政策與法規時的工具，建立國際間各國的互信互認（例如，國際認證與驗證機制、符合性評鑑機制）與 AI 共通語言，更有助於企業的國際化 AI 發展。

作者：梁日誠（FIAAIS｜AIMP｜CAIE、CCISA｜CCISM｜CISSP、GPM-b）現為 CMMC PI｜CCP｜CCA｜SME，加拿大 SCC／MC ISO／IEC JTC1／SC42、SC27、ISO／TC22／SC32、IEC／TC65 技術組成員，ISO 42001／ISO 27001／ISO 27701／ISO 22301／ISO 20000-1／IEC 62443-2-1 稽核師及講師，TCIC 環奧國際驗證公司全球營運總經理。