AI 的國際標準呼之欲出

ISO 42001即將公告 AI 管理有望獲得解方

AI 發展日新月異，如何放心使用已經成為當今研究的課題，尤其是如何確認 AI 的可信度，成為負責任的 AI，也成為重要關注的重要議題。

文／梁日誠

近年來，人工智慧（AI）應用推陳出新，大眾在看待 AI 新興科技的同時，也由不同的角度，如:安全、隱私、個資保護、法規遵循與風險等來檢視，以決定使用或接受 AI 與否，也因此，近來業界倡議「負責任的（Responsible）AI」，即期望 AI 的**可信度（Trustworthiness），包含如：透通性（Transparency）、可解釋性（Explainability）、可控制性（Controllability）、威脅與風險（Threats and Risks）、可用性（Availability）、韌性（Resiliency）、可靠度（Reliability）、正確性（Accuracy）、安全（Safety）、資安（Security）、隱私（Privacy）**等（可參考 2020 年公布的 ISO 24028 Overview of Trustworthiness in Artificial Intelligence）。

發展沿革

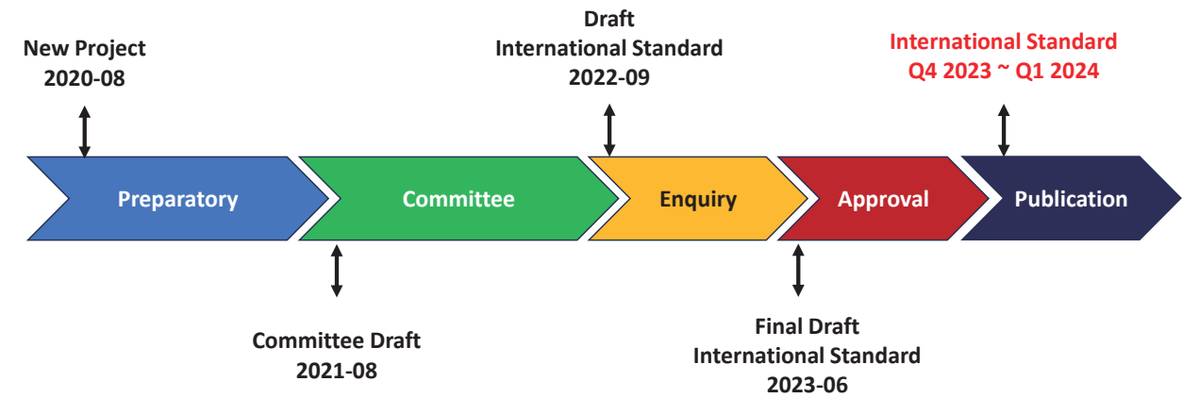
也因此，使用管理系統有系統地達成 AI 可信度的目標，循數個前例，如：資通安全的ISO 27001 資訊安全管理系統（ISMS）、隱私與個資保護的 ISO 27701 **隱私資訊管理系統（PIMS）**、業務持續與韌性的 ISO 22301 **業務持續管理系統（BCMS）**等，便應運而生了 **AIMS 人工智慧管理系統**，自 2020 年 8 月建立了 AIMS 國際標準 ISO／IEC 42001發展專案至今。

ISO 42001 自疫情漸緩後，發展進程（**如圖A**）也更積極，以支援國際間 AI 相關法規的進展，包括如：美國政府於 2020 年的第 13960 號總統令 Promoting the Use of Trustworthy AI in the Federal Government與AI in Government Act of 2020（DIVISION U, TITLE I）來管制聯邦政府的 AI 使用，NIST 也發布了Artificial Intelligence Risk Management Framework 標準來支持 AI 所需要的風險管理作業，值此之際，美國 AI 的法制化仍在持續進行中。

此外，以加拿大為例，Artificial Intelligence and Data Act（AIDA）正在法制化的過程中，AIDA 將要求企業由設計時期開始識別高衝擊的 AI 系統有關的傷害與偏差的風險，並於佈署時評鑑 AI 系統的預期使用及限制、確保使用者了解 AI 系統、施作適當的風險應對策略與確保 AI 系統的持續監控。同樣地，歐盟（EU）AI Act 也聚焦在高風險或高衝擊的AI系統。

適用範圍與目的

國際標準 ISO/IEC 42001（正發展中），由國際標準化組織（ISO）的 ISO/IEC JTC1/SC42 Artificial Intelligence 技術組主責，目前位於 FDIS（最終版國際標準草案）階段，**預計於 2023 年第三季至 2024 年第一季間公告**，此為國際標準（ISO/IEC）中針對 AI 的管理系統標準。



圖A

其中，ISO 42001 AIMS 旨在協助組織負責任地履行其在 AI 系統方面的角色（例如使用、開發、監控或提供利用 AI 的產品或服務），包括生成式（Generative） AI、機器學習（Machine Learning）、深度學習（Deep Learning）等。

ISO 42001的目的是指導組織如何管理（包括建立、部署、維護、持續改善），這即是專案管理中熟悉的Plan-Do-Check-Act（PDCA）過程，並提供要求（Requirements）。

在法規遵循的考量之下，管理系統被認為是針對合規（Compliance）所展現的有力方法，如同資訊安全管理系統 ISO 27001 對應於資通安全法規、隱私資訊管理系統 ISO 27701 對應於隱私與個資法規，人工智慧管理系統 ISO 42001 則對應於AI法規。

再者，ISO 42001 AIMS 並對組織的治理（Governance）、風險管理（Risk）及合規（Compliance）議題（簡稱 GRC），於 AI 領域提供了實踐的機制。

AI 的合規面

如同 ISMS／ISO 27001 包含資訊安全風險管理議題，PIMS／ISO 27701 包含隱私風險管理與隱私衝擊評鑑等議題；資訊安全風險管理可參考 ISO 31000 與 ISO 27005 標準，隱私風險管理與隱私衝擊評鑑可參考 ISO 31000 與 ISO 29134 等標準。對於AIMS／ISO 42001，則包含了 AI 風險管理與 AI 系統衝擊評鑑等議題。其中，AI 風險管理可參考 ISO 29384 標準（2023 年公告）、AI 系統衝擊評鑑可參考 ISO 42005 標準（目前正發展中）。

AI 相關標準共同對 AI 法規提供了有效的合規展現機制，以歐盟的 AI Act 為例，國際標準與其對應關係包含了對應於 Article 9 Risk Management System（與相關Articles）的 ISO 29384、對應於 Article 17 Quality Management System 的 ISO 42001 與對應於 Article 29 Obligations of Users of High-Risk AI systems 的 ISO 42005 及 ISO 29134（相關於 GDPR 的 Data Protection Impact Assessment），至於 AI 的概念與術語則列舉於 ISO 22989 標準（2022年公告）中供大眾檢閱。

ISO 42001 的內容概況

ISO 42001章節	章節名稱
第4節	組織全景
第5節	領導
第6節	規劃
第7節	支援
第8節	運作
第9節	績效評估
第10節	改善
附錄A	控制目標與控制措施
附錄B	AI控制措施實作指引
附錄C	潛在的AI相關的組織目標與風險來源
附錄D	跨領域或行業的AI管理系統使用

表A

各組織可以於其全景中識別出 AI 管理系統適用 的範圍，例如：以 AI 或隱私衝擊評鑑識別出涉及 AI 的相關系統或高風險或高衝擊的 AI 系統做為 AIMS 的範圍，對於 AIMS 範圍內的各系統進行 AI 風險評鑑，針對 AI 風險評鑑結果進行風險處理，包括選用附錄 A 的控制措施並參考附錄 B 的實作指引，持續進行績效評估、監控、量測控制措施有效性與風險管理，並針對不符合項與潛在不符合項持續改善 AIMS，來達到 Due Care 與 Due Diligence 的展現目的，持續地以負責任的方式使用 AI 系統。也如此，在 AI 的相關風險被控管的同時，才能安全地與有效地受惠於 AI 科技帶來的益處。

相關於風險的附錄 A 與附錄 B 的控制措施包含了 AI 相關的政策、內部組織、AI 系統的資源、評鑑 AI 系統衝擊、AI 系統生命週期、AI 系統的相關利害團體的資訊、AI 系統的使用、第三方關係等項目，並可依各組織的需要另增選控制措施。

至於，潛在的 AI 相關的組織目標與風險來源列舉於附錄 C，包含**表 B** 各項。附錄 D 則包含了「跨領域或行業的 AI 管理系統使用」的通則、AI 管理系統與其他管理系統標準的整合、驗證體制等相關項目。

組織目標（Objectives）	公平
	資安(Security)
	安全(Safety)
	隱私
	穩健
	透通性與可解釋性
	可歸責性
	可用性
	可維護性
	可用性與訓練資料品質
風險來源（Risk sources）	自動化等級
	缺乏透通性與可解釋性
	環境複雜度
	系統生命週期議題
	系統硬體議題
	科技準備度
	機器學習相關的風險來源

表B

可期待的負責任 AI

ISO 42001 涵蓋了 AIMS 要求事項，也因此，可以透過公正第三方的驗證來展現組織對 AI 的良善管理，帶給客戶、社會大眾、相關利害團體信心。至於目前正處於開發中（目前位於 DIS 國際標準草案階段）的 ISO 42006 則提供 AIMS 驗證與稽核機構的要求事項（即認證規範），於不久的將來，國際社會將可遵循完整的 AIMS 驗證與認證機制，並可與其高度相關的 ISMS、PIMS、BCMS 等充分整合成整合式管理系統（IMS），提供負責任的 AI 系統的良善管理與監督機制。

作者梁日誠（CISSP/CCISA/CCISM/GPM-b）現為 CMMC PI/CCP/CCA/SME，ISO/IEC JTC1/SC27、SC42、ISO/TC22/SC32、IEC/TC65 技術組委員，ISO 27001／ISO 27701／ISO 22301／ISO 20000-1／IEC 62443-2-1 主導稽核師及講師，TCIC 環奧國際驗證公司全球營運總經理。