英國發布《AI保證介紹》指引，藉由落實AI保證以降低AI系統使用風險

英國發布《AI保證介紹》指引，藉由落實AI保證以降低AI系統使用風險

資訊工業策進會科技法律研究所

2024年03月11日

人工智慧（AI）被稱作是第四次工業革命的核心，對於人們的生活形式和產業發展影響甚鉅。各國近年將AI列為重點發展的項目，陸續推動相關發展政策與規範，如歐盟《人工智慧法》（Artificial Intelligence Act, AI Act）、美國拜登總統簽署的第14110號行政命令「安全可靠且值得信賴的人工智慧開發暨使用」（Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence）、英國「支持創新的人工智慧監管政策白皮書」（A Pro-innovation Approach to AI Regulation）（下稱AI政策白皮書）等，各國期望發展新興技術的同時，亦能確保AI使用的安全性與公平性。

壹、事件摘要

英國科學、創新與技術部（Department for Science, Innovation and Technology，DSIT）於2024年2月12日發布《AI保證介紹》（Introduction to AI assurance）指引（下稱AI保證指引），AI保證係用於評測AI系統風險與可信度的措施，於該指引說明實施AI保證之範圍、原則與步驟，目的係為讓主管機關藉由落實AI保證，以降低AI系統使用之風險，並期望提高公眾對AI的信任。

AI保證指引係基於英國政府2023年3月發布之AI政策白皮書提出的五項跨部會AI原則所制定，五項原則分別為：安全、資安與穩健性（Safety, Security and Robustness）、適當的透明性與可解釋性（Appropriate Transparency and Explainability）、公平性（Fairness）、問責與治理（Accountability and Governance）以及可挑戰性 與補救措施（Contestability and Redress）。

貳、重點說明

AI保證指引內容包含：AI保證之適用範圍、AI保證的三大原則、執行AI保證的六項措施、評測標準以及建構AI保證的五個步驟，以下將重點介紹上開所列之規範內容：

一、**AI保證之適用範圍**：

（一）、**訓練資料（Training data）**：係指研發階段用於訓練AI的資料。

（二）、**AI模型（AI models）**：係指模型會透過輸入的資料來學習某些指令與功能，以幫助建構模模型分析、解釋、預測或制定決策的能力，例如GPT-4。，如GPT-4。

（三）、**AI系統（AI systems）**：係利用AI模型幫助、解決問題的產品、工具、應用程式或設備的系統，可包含單一模型或多個模型於一個系統中。例如ChatGPT為一個AI系統，其使用的AI模型為GPT-4。

（四）、**廣泛的AI使用（Broader operational context）**：係指AI系統於更為廣泛的領域或主管機關中部署、使用的情形。

二、**AI保證的三大原則**：鑒於AI系統的複雜性，須建立AI保證措施的原則與方法，以使其有效執行。

（一）、**衡量（Measure）**：收集AI系統運行的相關統計資料，包含AI系統於不同環境中的性能、功能及潛在風險影響的資訊；以及存取與AI系統設計、管理的相關文件，以確保AI保證的有效執行。

（二）、**評測（Evaluate）**：根據監管指引或國際標準，評測AI系統的風險與影響，找出AI系統的問題與漏洞。

（三）、**溝通（Communicate）**：建立溝通機制，以確保主管機關間之交流，包含調查報告、AI系統的相關資料，以及與公眾的意見徵集，並將上開資訊作為主管機關監理決策之參考依據。

三、**AI保證的六項措施**：主管機關可依循以下措施評測、衡量AI系統的性能與安全性，以及其是否符合法律規範。

（一）、**風險評估（Risk assessment）**：評測AI系統於研發與部署時的風險，包含偏見、資料保護和隱私風險、使用AI技術的風險，

以及是否影響主管機關聲譽等問題。

（二）、**演算法－影響評估（Algorithmic－impact assessment）**：用於預測AI系統、產品對於環境、人權、資料保護或其他結果更廣泛的影響。

（三）、**偏差審計（Bias audit）**：用於評估演算法系統的輸入和輸出，以評估輸入的資料、決策系統、指令或產出結果是否具有不公平偏差。

（四）、**合規性審計（Compliance audit）**：用於審查政策、法律及相關規定之遵循情形。

（五）、**合規性評估（Conformity assessment）**：用於評估AI系統或產品上市前的性能、安全性與風險。

（六）、**型式驗證（Formal verification）**：係指使用數學方法驗證AI系統是否滿足技術標準。

四、**評測標準**：以國際標準為基礎，建立、制定AI保證的共識與評測標準，評測標準應包含以下事項：

（一）、**基本原則與術語（Foundational and terminological）**：提供共享的詞彙、術語、描述與定義，以建立各界對AI之共識。

（二）、**介面與架構（Interface and architecture）**：定義系統之通用協調標準、格式，如互通性、基礎架構、資料管理之標準等。

（三）、**衡量與測試方式（Measurement and test methods）**：提供評測AI系統的方法與標準，如資安標準、安全性。

（四）、**流程、管理與治理（Process, management, and governance）**：制定明確之流程、規章與管理辦法等。

（五）、**產品及性能要求（Product and performance requirements）**：設定具體的技術標準，確保AI產品與服務係符合規範，並透過設立安全與性能標準，以達到保護消費者與使用者之目標。

五、**建構AI保證的步驟（Steps to build AI assurance）**

（一）、**考量現有的法律規範（Consider existing regulations）**：英國目前雖尚未針對AI制定的法律，但於AI研發、部署時仍會涉及相關法律，如英國《2018年資料保護法》（Data Protection Act 2018）等，故執行AI保證時應遵循、考量現有之法律規範。

（二）、**提升主管機關的知識技能（Upskill within your organisation）**：主管機關應積極了解AI系統的相關知識，並預測該機關未來業務的需求。

（三）、**檢視內部風險管理問題（Review internal governance and risk management）**：須適時的檢視主管機關內部的管理制度，機關於執行AI保證應以內部管理制度為基礎。

（四）、**尋求新的監管指引（Look out for new regulatory guidance）**：未來主管機關將制定具體的行業指引，並規範各領域實踐AI的原則與監管措施。

（五）、**考量並參與AI標準化（Consider involvement in AI standardisation）**：私人企業或主管機關應一同參與AI標準化的制定與協議，尤其中小企業，可與國際標準機構合作，並參訪AI標準中心（AI Standards Hubs），以取得、實施AI標準化的相關資訊與支援。

參、事件評析

AI保證指引係基於英國於2023年發布AI政策白皮書的五項跨部會原則所制定，冀望於主管機關落實AI保證，以降低AI系統使用之風險。AI保證係透過蒐集AI系統運行的相關資料，並根據國際標準與監管指引所制定之標準，以評測AI系統的安全性與其使用之相關影響風險。

隨著AI的快速進步及應用範疇持續擴大，於各領域皆日益重要，未來各國的不同領域之主管機關亦會持續制定、推出負責領域之AI相關政策框架與指引，引導各領域AI的開發、使用與佈署者能安全的使用AI。此外，應持續關注國際間推出的政策、指引或指引等，研析國際組織與各國的標準規範，借鏡國際間之推動作法，逐步建立我國的AI相關制度與規範，帶動我國智慧科技產業的穩定發展外，同時孕育AI新興產應用的發展並打造可信賴、安全的AI使用環境。

本文為「經濟部產業技術司科技專案成果」

→ 【高雄場】113年度新創採購-招標作業廠商說明會

劉心妍
副法律研究員 編譯整理

上稿時間：2024年06月

文章標籤

推薦文章