國家關鍵基礎設施應用人工智慧參考指引

114 年 10 月 8 日函頒

一、（適用對象及目的）

為利各領域國家關鍵基礎設施（以下稱 CI）之目的事業主管機關（以下稱主管機關）及所管領域之國家關鍵基礎設施提供者（以下稱設施提供者）於應用人工智慧（以下稱 AI）時，避免或緩解其影響網路安全、資料保護、營運持續及法規遵循等風險，爰訂定本指引。

二、（盤點並遵守相關法令）

主管機關及設施提供者於 CI 應用 AI 時，應盤點並遵守相關法令規定，包括個人資料保護法、資通安全管理法、政府資訊公開法、著作權法及各領域主管機關訂頒之相關作業規定等[1]。

三、（主管機關應辦事項）

主管機關應就所管領域設施特性，及時調適所管相關法令；依據「國家關鍵基礎設施安全防護指導綱領」辦理所管領域設施安全防護事宜時，亦應就所管領域設施特性，將應用 AI 風險評估及管理措施納入考量，從領域層級高度協助設施提供者建立相應之風險評估及管理能量，包括：撰擬之「領域安全防護計畫」應涵蓋應用 AI 風險評估及管理；專案團隊清點轄下可能重要資產及設施時，應併同盤點各設施應用 AI 情形；辦理領域層級資訊分享及教育訓練時，應涵蓋國內外應用 AI 進展、風險及管理對策。

四、（設施提供者應辦事項）

設施提供者依據「國家關鍵基礎設施安全防護指導綱領」辦理設施安全防護事宜時，應符合主管機關相關規定及領域安全防護計畫，並就該設施特性，將應用 AI 風險評估及管理措施納入考量，包括：撰擬之「國家關鍵基礎設施安全防護計畫」應涵蓋應用 AI

[1] 參考數位發展部所訂定「公部門人工智慧應用參考手冊」之「AI 應用涉及法規指引」一節。

風險威脅辨識、評估及管理，並定期更新；指定之專責組織與人員推動及監督安全防護相關事務時，應定期盤點該設施中所有應用 AI 情形，建立清單並定期更新；召開安全防護會議時，應檢核風險評估及管理措施執行成效並改進。

五、（CI 應用 AI 風險類型）

主管機關及設施提供者識別 AI 風險類型，得依領域及設施特性，參考下列七大類風險辦理：

（一）AI 系統本身之安全性、故障及限制。

（二）人機互動失衡。

（三）隱私及智慧財產權侵害。

（四）歧視及生成違法內容。

（五）錯誤訊息及惡意詐騙。

（六）惡意行為及攻擊。

（七）社會經濟環境危害。

主管機關及設施提供者除依前項規定辦理外，亦得依領域及設施特性，參考先進國家分類作法，識別應用 AI 風險類型，例如參考美國建議如下[2]：

（一）利用 AI 進行之攻擊。

（二）針對 AI 系統之攻擊。

（三）AI 設計及實施失敗。

六、（AI 風險程度及衝擊評估）

設施提供者應評估 AI 風險程度及衝擊影響，並就高風險或高衝擊 AI 應用，採行事前許可、審查等措施緩解風險；無手段可管理或降低風險及衝擊時，應予以限制或禁止。

七、（事前、事中及事後風險管理措施）

設施提供者依資通安全管理法訂定、修正及實施資通安全維護計

[2] 參考美國國土安全部「緩解人工智慧風險：關鍵基礎設施提供者的保護與安全指引」（MITIGATING ARTIFICIAL INTELLIGENCE (AI) RISK: Safety and Security Guidelines for Critical Infrastructure Owners and Operators. April, 2024.）

畫時，應依前二點風險識別及評估結果，依設施特性參考採行下列防護及控制措施，於資安事件發生前、中、後各階段，緩解應用 AI 風險及衝擊影響：

（一）事前預防

1. 建立韌性：制定及實施風險管理計畫，以確保事件發生時仍能維持營運，例如：備援（份）系統、營運持續計畫及演練等。
2. 納入安全設計原則：產品之安全性被視為核心業務要求，並於產品開發生命周期之設計階段即納入考量。
3. 建立軟體物料清單：詳細列出軟體之所有元件、依賴關係及其結構，並附上易受攻擊元件之識別及交換資訊。
4. 採用存取控制：限制對訓練資料、模型及輸入之存取，採用強密碼、多因子驗證及最小權限原則等措施控管存取權限。評估是否限制其資料儲存之所在地為我國管轄權所及之境內，並依存放於雲端資料之重要性，評估是否需對資料進行加密。
5. 使用加密機制：使用公開、國際機構驗證且未遭破解之演算法進行加密。
6. 採用內容識別技術：採用浮水印或驗證技術，協助辨識 AI 生成之媒體內容，減少深偽及假資訊造成之欺詐。
7. 防護網路安全：保護網路基礎設施免於未經授權存取、濫用或竊取，例如：存取控制或網路區隔等。
8. 執行漏洞管理：主動識別、通報與修補 AI 模型及其相關軟體中之漏洞。
9. 納入人為監督：於 AI 之開發、部署及操作過程中納入人為監督，以提升問責性。
10. 強化資安意識：透過資安基礎及實務訓練，教導使用者基本之安全行為，以防範網路威脅。

（二）事中應處

1. 進行事件通報：依「資通安全事件通報及應變辦法」辦理，知悉資安事件後，於一小時內依主管機關指定之方式進行通報。
2. 組成應變小組：召開事件應變會議，了解事件概況及評估受影響範圍。
3. 執行損害控制及復原：依「資通安全事件通報及應變辦法」規定時間完成損害控制或復原作業[3]，確認具體受害範圍，執行營運持續計畫，優先恢復對外服務及核心資通系統運作，例如：啟用備份資料或啟動備援機制；備援（份）系統包括人工或非人工智慧系統，以應對中斷期間之營運需求。
4. 進行根因分析：得請專業廠商或專家進行檢測及分析事件根因。

（三）事後追蹤

1. 進行情資分享：政府機關與設施提供者間之威脅情資交流，包括實體與網路安全威脅資訊及防禦經驗。
2. 驗證資料模型及功能：驗證AI模型是否如預期運作，例如：已驗證之測試資料集、第三方審查或外部認證機構。
3. 追蹤事件改善：評估改善作為期程及改善策略，送交調查、處理及改善報告予主管機關。
4. 進行跡證保存：進行跡證保存時，應優先採取隔離機制，備份受害系統儲存媒介，例如：硬碟或虛擬機映像檔。

[3] 第一級或第二級資通安全事件，於知悉該事件後七十二小時內。第三級或第四級資通安全事件，於知悉該事件後三十六小時內。