國家資通安全研究院
National Institute of Cyber Security

資安週報

Cyber Security Weekly Newsletter

事件通報
帳號密碼如涉外洩 宜儘速完成憑證更換並檢視相關登入情形

聯防監控
防禦迴避高居首位 偵測刺探仍具威脅

蜜罐誘捕
企業級網路防火牆軟體成攻擊熱點

重大漏洞
Cisco、Apache及Oracle高風險漏洞應盡速修補

外部曝險分析
元件高風險漏洞大幅增加 為本期外部曝險風險上升主因

實兵演練
實兵演練揭露遠端管理設備漏洞及預設帳號密碼使用風險

網路巡查高風險詐騙
免費評估與限時優惠推升詐騙風險

焦點文章
115年上半年國際AI安全治理政策發展脈動

2026.07.02
051

資安儀表板

事件通報、聯防監控、蜜罐誘捕、重大漏洞、外部曝險分析及實兵演練等六類量化指標

■事件通報

近一週公務機關資安事件通報之類型與數量，同時包含民營機構依規定揭露重大資通安全訊息

帳號密碼如涉外洩 宜儘速完成憑證更換並檢視相關登入情形

本週總計接獲16件公務機關與特定非公務機關事件通報，詳見圖1，公務機關非法入侵事件中以異常連線占多數，詳見圖2。本週有機關通報該機關因應FortiBleed事件，經查詢驗證發現外流資料中包含曾使用過之帳號密碼。依Fortinet公開說明，該事件主要係涉及過往外洩資料再次流傳，並非新的Fortinet漏洞；惟若相關管理員或 VPN 帳號仍沿用外洩憑證、密碼強度不足，或未啟用多因素驗證，仍可能增加設備遭未授權存取之風險。

建議如發現使用中帳號密碼涉及外洩資料，應立即中止相關管理員及VPN連線、重設帳號密碼並強制啟用多因素驗證，同時檢查設備設定與日誌，留意是否存在異常管理登入、未授權帳號、可疑設定異動或非預期VPN連線情形；另宜導入白名單機制，限制僅供特定來源存取，降低管理介面直接暴露於網際網路之風險。

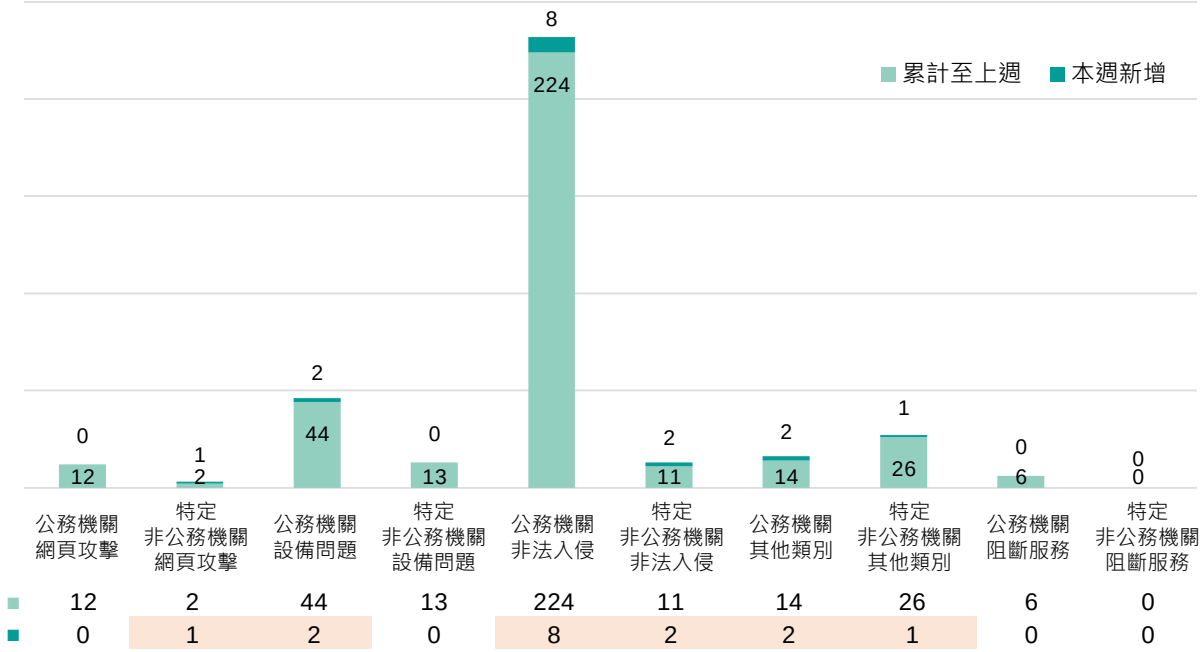

圖1 | 本週公務機關暨特定非公務機關資安事件通報概況

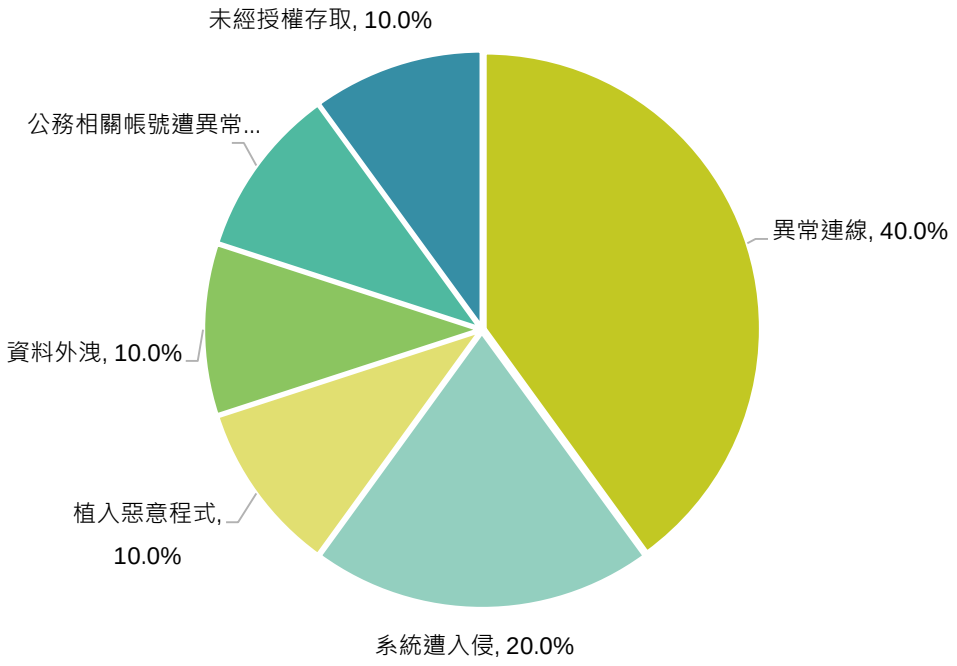

圖2 | 本週公務機關非法入侵事件類型占比

防護建議 **除修補漏洞外，應：**

以攻擊為出發評估潛在風險	針對潛在風險執行相應改善
➢ 歷史外洩憑證遭惡意利用，沿用舊密碼之管理與VPN帳號易遭破入 ➢ 設備若缺乏多因子驗證，駭客登入後可繞過偵測並向內網滲透	➢ 發現異常應立即中止涉嫌外洩之連線，並強制要求相關管理員變更密碼 ➢ 全面針對管理與VPN帳號啟用多因子驗證(MFA)，防範憑證遭竊取冒用 ➢ 深入審查設備設定與操作日誌，嚴防未授權帳號建立或非預期VPN連線 ➢ 導入來源IP白名單機制，嚴禁管理介面直接暴露於網際網路，降低受攻擊面

3間民間企業揭露重大資安訊息

本週3家民間企業發布重大訊息，產業類別為電機機械業、其他電子業、電腦及週邊設備業，詳見表1。

表1 | 本週民間企業重大資安訊息彙整

公司名稱	發布時間	事件說明
易發精機股份有限公司	115年6月22日	易發之子公司東野精機(股)公司日福精工(股)公司，發現遭不明駭客入侵主機，並將伺服器資料加密，導致部分系統無法使用。 資安部門立即啟動相關資安防禦與系統復原，並委請外部資安技術公司及專家共同處理。 目前初步評估對公司運作無重大影響，後續將持續提升網路與資訊基礎架構之安全管控，以確保資訊安全。
耕興股份有限公司	115年6月22日	耕興部分資訊系統發生網路資安事件，察覺異常時已立即啟動資安防禦、復原機制並委請外部資安公司技術專家共同處理，後續已依程序報警。 受影響之資料檔案留有備份，不影響公司營運、業務進行等，後續將持續提升網路與資訊基礎架構之安全管控，並依外部資安技術專家之建議加強防護。
尼得科超眾科技股份有限公司	115年6月22日	尼得科超眾偵測到部分伺服器遭受網路勒索病毒攻擊，致使包含ERP在內之部分資訊系統無法正常運作，資安團隊於第一時間已立即啟動防禦機制與斷網隔離措施。 對公司整體營運及財務之影響程度尚在評估中，將視系統復原進度，適時調整生產與出貨排程。

■聯防監控

近一週以MITRE ATT&CK Matrix 分析攻擊者行為，提醒公務機關留意攻擊趨勢變化，是否由初期之偵測刺探進入影響層面更大竊取資料與破壞資通系統

防禦迴避高居首位 偵測刺探仍具威脅

本週政府領域資安聯防監控參考MITRE ATT&CK Matrix分析TTP戰術框架分布顯示，本週趨勢相較上週無顯著差異，詳見圖3。「防禦迴避」為最常見攻擊手法，占比15.9%，攻擊者通常會透過關閉或刪除指令紀錄，並利用合法的系統工具間接執行惡意命令，以達到規避監控的目的。因應此類威脅，建議導入端點防護措施，加強指令紀錄的稽核能力，限制高風險工具的濫用，並強化特權帳號的管理，以避免攻擊者繞過偵測並消除其行為痕跡。

「偵測刺探」事件本週占比為13.8%，為本週占比次高的攻擊階段，顯示攻擊者持續加強對目標環境的前期偵查與情報蒐集行動，以利後續攻擊規劃與目標鎖定。觀察到的主要手法包括IP 區段掃描、主動式掃描、以及憑證資訊蒐集。攻擊者透過自動化工具對外部網段與公開服務進行大規模探測，識別可存取主機、開放埠及系統服務，並蒐集與帳號憑證相關的資訊，作為後續登入攻擊或權限取得的基礎。此類活動通常由廣泛掃描逐步聚焦至特定目標，具備明確的階段性與目的性，能有效提升後續攻擊的成功率。建議持續監控異常掃描流量與帳號探測行為，定期盤點對外暴露資產，並結合威脅情資分析可疑來源，以提前辨識攻擊者的偵查活動並及早採取防禦措施。

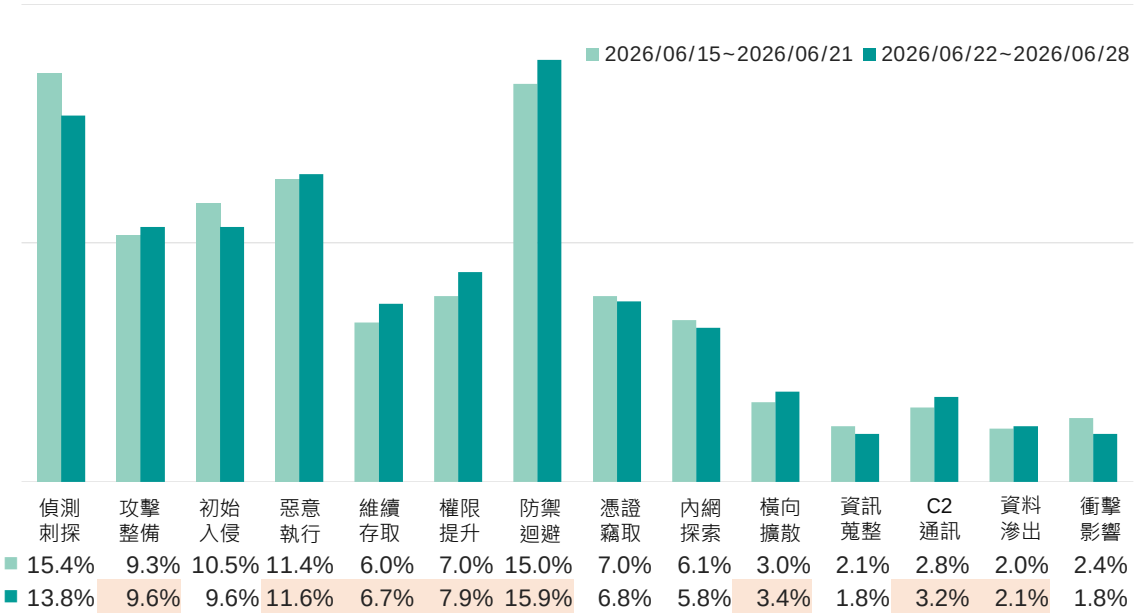

圖3 | 資安聯防監控攻擊階段統計

防護建議 **建議機關採取下列防護措施：**

防禦迴避(Defense Evasion)

- 啟用端點防護(EDR)及行為監控機制
- 保留並定期稽核系統與指令執行紀錄（如 PowerShell、Shell、Windows Event Log）
- 限制高風險系統工具(Living-off-the-Land Binaries, LOLBins)使用權限
- 落實最小權限原則，加強特權帳號管理與多因素驗證(MFA)
- 建立異常行為告警機制，偵測日誌清除、服務停用等可疑活動

偵測刺探(Reconnaissance)

- 持續監控異常掃描流量及帳號探測行為
- 定期盤點並降低對外暴露資產與服務
- 關閉非必要連接埠及未使用的網路服務
- 部署入侵偵測/防禦系統(IDS/IPS)及網路流量分析機制
- 結合威脅情資(Threat Intelligence)封鎖惡意來源IP，並強化外部資產監測

■蜜罐誘捕

近一週誘捕系統所捕捉到的攻擊樣態趨勢變化以及所利用的弱點趨勢

企業級網路防火牆軟體成攻擊熱點

本週透過部署於國內外之蜜罐系統觀測攻擊行為動態，相較於上週「網頁應用」服務攻擊占比71.71%、「遠端控制」服務攻擊占比24.25%，本週各類服務之平均偵測攻擊比例無明顯變化，結果顯示「網頁應用」服務仍為攻擊主軸，占比高達80.37%。「遠端控制」服務亦有16.64% 的誘捕比例，反映攻擊者仍積極針對公開遠端連線介面進行入侵行動。

網頁應用是最為常見之對外服務類型，若存在已知漏洞，將面臨高風險曝露情形，易成為攻擊者入侵與滲透重要管道，為優先防護之項目。本週網頁應用介面之誘捕狀況，詳見圖4。本週通用型Web介面占比最高，此類別為攻擊者廣泛的進行HTTP掃描與探測，顯示攻擊者企圖尋找可能存在之Web漏洞進行攻擊。

另Web服務系統類別包含各類以網頁為基礎的服務與應用，例如常見的網頁框架、應用程式伺服器、檔案傳輸與資料管理平台等，由於此類服務多建置於企業應用環境，且直接面向外部提供功能與資料交換，為僅次於通用型介面的攻擊目標。而Web服務系統比例下降主因為cPanel & WHM在登入流程中存在身份驗證繞過漏洞的CVE-2026-41940，遭攻擊次數下降導致。網通設備管理介面涵蓋路由器、防火牆等網通設備管理介面，以及智慧攝影機、NAS等物聯網設備管理介面，皆容易遭受攻擊者透過弱密碼、預設帳號或已知漏洞進行入侵。其他類型服務雖比例極小，但若涉及關鍵業務系統，仍需留意潛在風險。

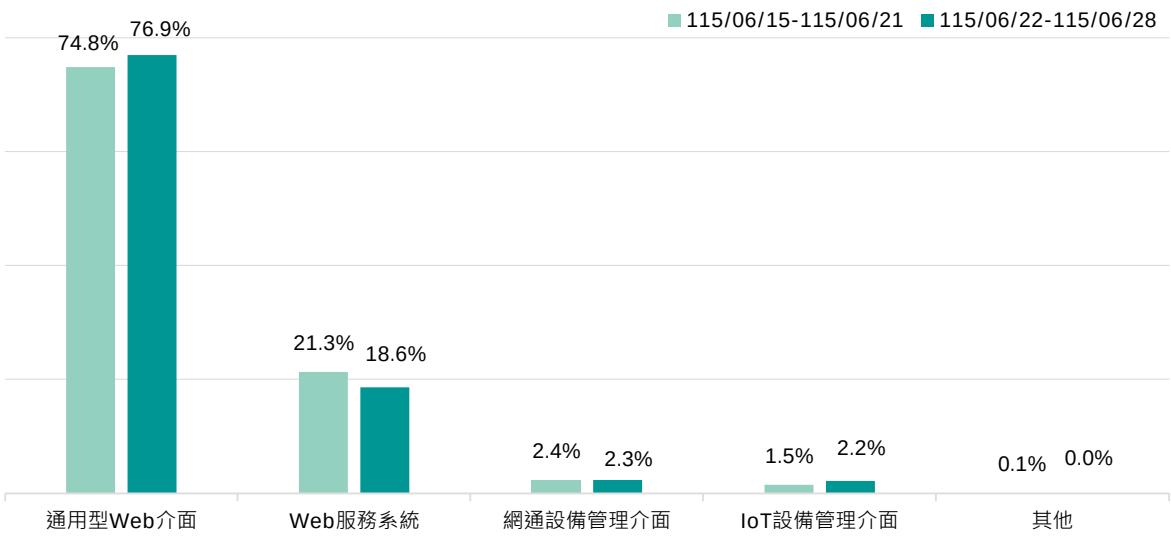

圖4 | 本週網頁應用介面之誘捕攻擊比例統計

進一步解析國內外之蜜罐系統誘捕漏洞攻擊之情形，詳見表2。近2年揭露之攻擊漏洞，前5大攻擊以「網頁應用」服務之漏洞為主要入侵路徑，本週漏洞類型多集中於授權缺失漏洞、越界讀取漏洞、跨站請求偽造、資訊外洩漏洞及路徑遍歷漏洞，攻擊目標涵蓋企業級網路防火牆軟體、程式遞送控制器(ADC)、多選下拉選單元件、開源後端伺服器框架及內容管理系統之網站優化外掛程式。

防護建議

建議存在漏洞之設備應更新至最新版本軟體或韌體以修補漏洞；若原廠已無法提供更新支援，應考慮汰換存在漏洞之設備或軟體套件，如因故無法汰換，應採對應之漏洞緩解措施。

表2 | 本週前5大攻擊使用之近2年漏洞排行列表

排名			漏洞編號	受影響產品	CVSS 3.x Base Score
■	1	⬆1	CVE-2025-20362[1]	Cisco Secure ASA & FTD	6.5
■	2	⬇1	CVE-2025-5777[2]	Citrix NetScaler ADC	7.5
■	3	-	CVE-2025-47204[3]	Bootstrap Multiselect	6.1
■	4	⬆new	CVE-2025-53364[4]	Parse Server	5.3
■	5	-	CVE-2025-30567[5]	WordPress WP01	7.5

類型 ■ 授權缺失漏洞 ■ 越界讀取漏洞 ■ 跨站請求偽造 ■ 資訊外洩漏洞 ■ 路徑遍歷漏洞

1. https://nvd.nist.gov/vuln/detail/CVE-2025-20362
2. https://nvd.nist.gov/vuln/detail/CVE-2025-5777
3. https://nvd.nist.gov/vuln/detail/CVE-2025-47204
4. https://nvd.nist.gov/vuln/detail/CVE-2025-53364
5. https://nvd.nist.gov/vuln/detail/CVE-2025-30567

■重大漏洞

近一週本院研究人員發現以下重大漏洞資訊，建議組織內部進行檢查與修補

Cisco、Apache及Oracle高風險漏洞應盡速修補

- Cisco Identity Services Engine(ISE)與ISE Passive Identity Connector(ISE-PIC)存在遠端程式碼執行(Remote Code Execution)漏洞(CVE-2026-20181[1])，已取得管理權限之遠端攻擊者可藉由發送特製HTTP(S)請求，於設備上以root權限執行任意程式碼。

- Apache HTTP Server存在3個高風險安全漏洞(CVE-2026-23918[2]、CVE-2026-29167[3]及CVE-2026-44631[4])，類型包含記憶體雙重釋放(Double Free)、使用釋放後記憶體(Use After Free)及緩衝區溢位(Buffer Overflow)漏洞，其最嚴重可使已通過身分鑑別之遠端攻擊者執行任意程式碼。

- Oracle釋出115年6月安全性更新，已修補PeopleSoft Enterprise PT PeopleTools、WebLogic Server及Identity Manager等共244個漏洞[5]，其中包含159個高風險漏洞，另CVE-2026-35273[6]已遭駭客利用。

1. https://nvd.nist.gov/vuln/detail/CVE-2026-20181
2. https://nvd.nist.gov/vuln/detail/CVE-2026-23918
3. https://nvd.nist.gov/vuln/detail/CVE-2026-29167
4. https://nvd.nist.gov/vuln/detail/CVE-2026-44631
5. https://www.oracle.com/security-alerts/cspujun2026.html
6. https://nvd.nist.gov/vuln/detail/CVE-2026-35273

■外部曝險分析

經由外部檢測政府機關資通安全狀況，例如使用EASM工具或實兵演練，及早發現曝露於外部之風險

元件高風險漏洞大幅增加 為本期外部曝險風險上升主因

本次針對曝險程度較高之100個A/B級公務機關進行EASM資安曝險檢測，前10大風險項目共計10,636項。其中，「元件高風險漏洞」以5,008項居首，「過時或弱加密協定」2,016項次之，「TLS憑證不受信任」1,531項位居第三。前三項合計8,555項，占前10大風險項目總數約80.4%，顯示目前外部曝險風險仍高度集中於元件漏洞、加密協定與憑證管理等議題。相較上期9,783項，本期整體風險數量增加853項，增幅約8.7%，詳見圖5。

進一步分析主要風險變化情形，「元件高風險漏洞」仍為最大宗風險，數量由上期4,192項增至本期5,008項，增加816項，增幅約19.5%，為本期整體風險上升之主要原因。其中，Apache HTTP Server相關弱點(CVE-2024-38475)仍較為集中。「過時或弱加密協定」由1,959項增至2,016項，增加57項，增幅約2.9%；「TLS憑證不受信任」由1,632項減至1,531項，減少101項，降幅約6.2%。另「CSP設定不當」由888項增至924項，增加36項，增幅約4.1%；「未部署WAF」由684項增至759項，增加75項，增幅約11.0%。整體而言，本期外部曝險風險增加主要來自元件高風險漏洞大幅成長，後續宜優先追蹤相關系統、元件及第三方套件之版本更新與修補情形。

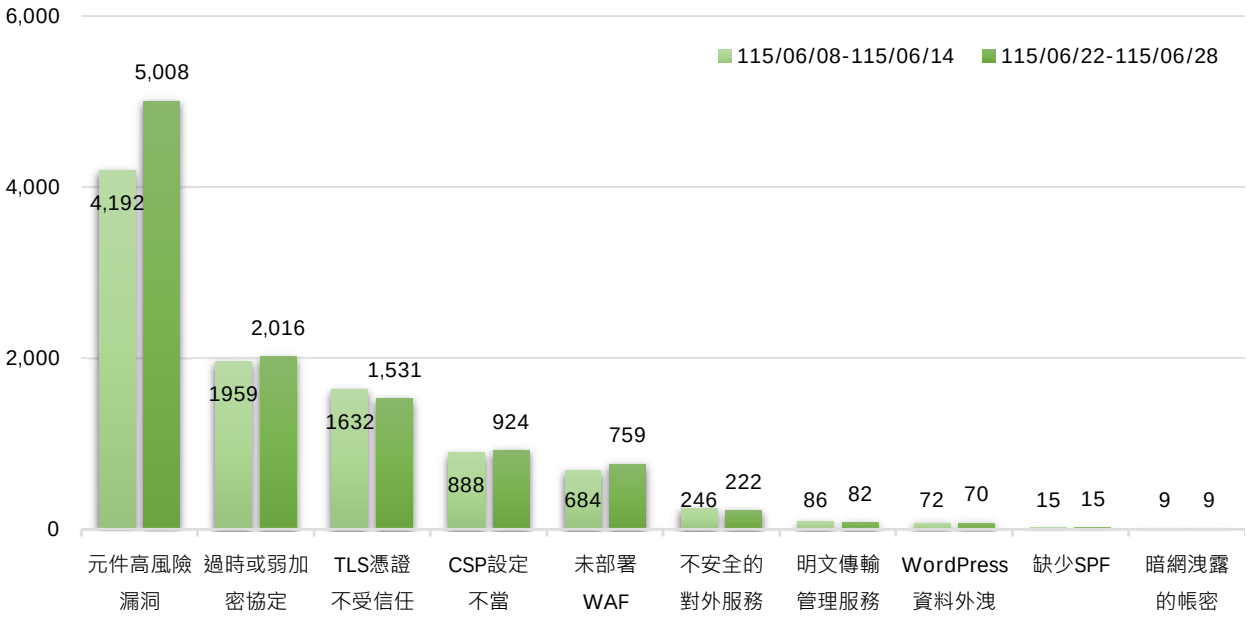

圖5 | EASM檢測結果統計(前10大風險)

防護建議

建議機關或關鍵基礎設施採取下列防護措施

- 定期更新憑證，全面啟用TLS 1.2以上版本協定，並停用未加密或舊版協定
- 儘速修補已知漏洞，並淘汰已無維護之軟體版本
- 部署網站應用程式防火牆(WAF)，並導入內容安全政策(CSP)等網站安全標頭，以降低XSS攻擊與惡意存取風險
- 關閉不必要之對外服務，並定期盤點已停用站台、主機與應用服務是否已確實完成下線，避免因資產未完全關閉而持續對外曝露；如確有遠端管理需求，應嚴格限制來源IP，並採用加密通道(如SSH)

建議機關或關鍵基礎設施採取下列管理措施

- 定期盤點並更換外洩帳號憑證，搭配多因素驗證(MFA)，以強化存取安全
- 建立弱點修補與驗證機制，確保風險持續改善
- 強化資安教育訓練，提升系統維運人員對憑證管理、加密配置及服務設定之安全意識

■實兵演練

實兵演練係模擬真實攻擊情境，檢驗資通系統弱點、防護能力及應變機制，以提升整體資安防禦韌性

實兵演練揭露遠端管理設備漏洞及預設帳號密碼使用風險

本年資通系統實兵演練針對政府機關與關鍵基礎設施提供者，於擇定時間進行演練，至6/26止已累計188筆攻擊紀錄，本週新增弱點數量共19筆，分別為「不安全的組態設定」7筆、「驗證機制失效」4筆、「無效的存取控管」3筆、「注入攻擊」2筆、「軟體供應鏈失效」2筆及「加密機制失效」1筆，詳見圖6。

本週演練發現，衝擊性較高之弱點為「軟體供應鏈失效」。攻擊者於遠端管理設備登入介面取得設備版本資訊後，確認該弱點CVSS v3.1分數為10.0(Critical)，透過公開漏洞利用資料庫取得對應之漏洞利用程式(Exploit)，成功利用漏洞新增使用者帳號，並以該帳號登入管理介面。另發現部分IoT設備系統存在「驗證機制失效」弱點，攻擊者利用常見帳號密碼成功登入設備管理介面，接續於掃描發現之路徑取得相關系統檔案，並由檔案內容發現其他系統路徑，經瀏覽取得系統管理者明文帳號密碼，進而使用該組帳號密碼成功登入系統。

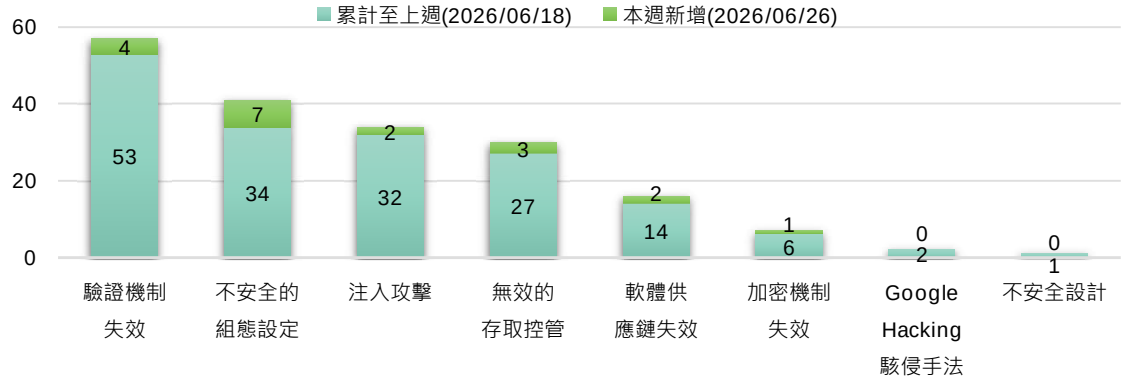

圖6 | 網路攻防演練資通系統實兵演練統計

建議機關及關鍵基礎設施提供者，採取下列防護措施

- 管理介面應限制來源IP或管理網段存取，避免直接暴露於外部網路；如因業務需求須對外提供服務，應採取VPN、多因素驗證(MFA)等安全機制，以降低遭未授權存取之風險
- 應定期盤點IoT設備與遠端管理設備之韌體版本，持續追蹤原廠安全公告與CVE漏洞資訊，並儘速完成安全修補，避免已知漏洞遭利用
- IoT設備不建議使用預設或常見帳號密碼(如admin/admin)，應建立高強度密碼政策，並定期更換管理者密碼
- 系統管理者帳號密碼與其他敏感資訊應妥善管理，避免以明文方式儲存於系統檔案、設定檔或網頁中，並依最小權限原則限制存取權限

■網路巡查高風險詐騙

追蹤詐騙訊息與手法演變，掌握政府機關實施之打詐政策與機制，是否達成其控制目標

免費評估與限時優惠推升詐騙風險

本週高風險詐騙關鍵字分析

本週高風險廣告以「優惠」出現次數最高，其次為「健康」、「投資」、「配方」、「免費」、「台灣」、「AI」、「推薦」、「翡翠」及「日本」。相關訊息多以投資課程、保健食品、護眼產品、AI創業課程、二手車貸、珠寶競拍及收藏品拍賣吸引民眾，再以限時折扣、免費評估、醫師推薦、官方正品、私訊加 LINE 或點擊連結導流，詳見圖7。

「優惠、免費」降低警覺

「優惠」與「免費」常見於投資課程、期貨入門、貸款方案及副業招募，搭配「低於 5 折」、「留言領優惠券」、「限時 82 折」、「免費報名」、「免費評估」等說法，誘導民眾填表、加 LINE、點擊連結或付款。提醒民眾，低價與免費不代表可信，仍應查證業者身分、服務內容與退費規範。

圖7 | 詐騙關鍵字文字雲

「投資、股票、AI」包裝賺錢機會

本週多則廣告以美股、台股、ETF、期貨、AI創業及財務諮詢課程為主，宣稱可建立投資策略、使用AI分析工具或創造第二收入。提醒民眾，投資與副業均有風險，任何課程、工具或系統都不能保證獲利；報名前應確認公司資訊、合約內容及風險揭露。

「健康、配方、醫師推薦」鎖定身體焦慮

保健食品、護眼液、眼霜、草本茶飲及外用產品常以「調整體質」、「三高調理」、「無需手術」、「快速改善」、「3 天消眼袋」、「讓頭髮黑起來」等話術吸引購買。凡商品宣稱可改善疾病、調節血糖血壓血脂或取代醫療行為，均應查證合法標示與產品來源。

「台灣、日本、正品、翡翠」營造可信感

護眼產品標榜醫師推薦、日本漢方；珠寶與收藏品則強調日本回流、GIA 證書、二手正品、假一賠十、競拍撿漏。提醒民眾，國家來源、證書或正品保證不代表交易安全，涉及翡翠、鑽石手鏈、銀壺等高價商品時，應確認賣家身分、鑑定資料與交易保障。

「私訊、點擊、下單」是導流警訊

中古車貸、工作燈、手鏈、AI創業、財務諮詢及投資課程廣告，常要求加 LINE、私訊、填寫試算表或點擊連結報名。凡價格、貸款條件、課程內容、測算結果或下單方式必須私下提供者，應暫停操作並查證來源。

本週防詐提醒

1. 「優惠」、「免費」、「限時」不代表零風險。
2. 投資、股票、期貨、AI工具及財務諮詢課程不得保證獲利。
3. 保健與草本商品若宣稱快速改善疾病，應查證合法標示。
4. 翡翠、鑽石手鏈、日本銀壺等高價商品，應確認鑑定與交易保障。
5. 不要將身分證、金融帳戶、信用卡、驗證碼或病歷提供給陌生賣家、客服、貸款人員或課程招募者。

綜合觀察

本週高風險廣告多以「優惠」與「免費」作為入口，再透過「投資」、「健康」、「AI」放大需求，並以「台灣」、「推薦」、「日本」、「正品」包裝可信度，最後導向私訊、LINE、表單或付款流程。遇到來源不明、資訊不完整、要求提供個資或先付款的廣告，應先查證再決定。

焦點文章

115年上半年國際AI安全治理政策發展脈動

強調主權與治理，代理AI成為各國當前焦點

隨著生成式 AI(Generative AI)迅速發展，國際間AI治理的問題意識，已從「要不要管、訂什麼原則」，進一步推進到「用什麼制度去管、由誰來確認安全」等議題。本文從政策治理的角度，綜整2026年上半年國際間與AI安全(Safety)及資安(Security)相關的動態，梳理這段期間的發展趨勢。

當前值得注意的趨勢有三：各國管制路線雖有不同，但均重視主權；治理工具尋求落實；以及代理AI(Agentic AI)成為治理新標的。文末並就這些趨勢，提出對我國可能的政策建議。本文參考資料取自本院《國際資安政策法制觀測》[1]週報對各國的長期追蹤。

一、各國的管制路線雖有不同但均重視主權

首先，美國川普政府在近期發佈《促進先進人工智慧創新與安全》行政命令[2]，採取「自願協作而非強制授權」的路線，不另設強制授權或預先審查機制，而把AI風險治理與國家安全、關鍵基礎設施防護結合，並以聯邦既有機關與自律標準替代繁重管制。這條去管制而重國安的路線，是川普在第二任期當中對AI的政策基調。

歐盟同樣以自主為政策重點，作法則是雙軌並進。一方面以數位綜合法案(Digital Omnibus)修法提案簡化《人工智慧法》(AI Act)的合規負擔，提供中小企業比例性的合規措施[3]；另一方面推出歐洲科技主權套案(European Technological Sovereignty Package)[4]，透過相關立法設定五至七年內把資料中心容量增為三倍的目標。簡化合規與強化主權同時推進，顯示歐盟一面維持規範密度，一面正視過度依賴境外供應商的風險。

亞洲主要國家則走第三條路，把治理嵌進法律與行政流程。韓國的《人工智慧基本法》於今年一月施行，被視為繼歐盟之後全球第二部綜合性AI法律，課予業者高影響AI判定、生成內容透明度標示等義務，並設約一年的緩衝期，整體採「促進與信任並重」的取向[5]。日本則設置首席 AI官(Chief AI Officer, CAIO)統籌各機關的AI治理，並透過政府採購指引，把風險評估與責任分工嵌入行政流程[6]。

三種路線的共同點是主權自主：無論透過國安、立法或算力布局，各國都在爭取對AI基礎設施、標準與評估能力的自主掌握。

焦點文章

二、治理工具尋求落實

第二個趨勢是，治理工具正從抽象原則轉為可操作、可落實的制度。過去兩年各國競相發布AI原則與倫理宣言，近期則進入落地階段。韓國在基本法施行後，發布支援窗口案例集，就「高影響AI」判定、透明度標示等義務提供企業可遵循的操作解答[7]；並於五月預告施行令修正案，銜接七月生效的修正後基本法[8]。日本數位廳的《生成式AI採購與利用指引》2.0版[9]，則以風險評估與CAIO三層治理機制，把生成式AI的採購與高風險用例管理系統化。政策的著力點，已從「宣示要負責任」轉向「規定怎麼做才算負責任」。

落地的另一個面向，是量測能力的強化，各國陸續設立的AI安全研究機構(AI Safety / Security Institute，簡稱AISI)，正成為研究AI驗證與測試方法的核心單位。英國AI安全研究院(AI Security Institute)發布多步驟網路攻擊的評測研究，以受控靶場量測模型在不同推論算力下完成攻擊任務的能力[10]；美國NIST人工智慧標準暨創新中心(Center for AI Standards and Innovation, CAISI)等則以大規模公開紅隊競賽檢驗代理(Agent)的脆弱性，並與多家前沿實驗室簽訂部署前的國安測試協議[11]。日本人工智慧安全研究所(J-AISI)與資訊處理推進機構(IPA)合辦的JAI-Trust計畫，設置十一個分科委員會，分別針對偏見、越獄、資訊洩漏、代理型AI、多模態等風險建立可共用的評估基準[12]。這些機構並開始相互協作，例如英國與澳洲(Australian AI Safety Institute)於五月簽署合作備忘錄，共享前沿模型的能力評估與最佳實務[13]。

制度落地也改變了「安全」的定義。NIST 研究員證明，任何有限的護欄(guardrails)規則集都必然存在可規避它的對抗性提示，這是數學定理的延伸結果，而非實作缺陷[14]。換言之，要求業者建立「完整防護」形同要求達成不可能之事。資安政策的重點，因而從追求完整防護，移向強化監控覆蓋與應變速度：NIST主張為已部署系統建立「部署後監控、回饋、再設計」的循環，英國AISI的「喪失監管」(loss of oversight)報告也提醒，監管能力需要從設計、訓練到部署各階段主動建置[15]。

三、代理型AI成為治理新標的

第三個趨勢，是治理對象由「模型的輸出內容」推進到「能自主行動的系統」。具自主規劃與工具調用能力的代理型AI，已成為各國治理框架的新焦點。

焦點文章

新加坡率先發布全球首份由政府發布、代理專屬的《AI代理治理框架模型》(Model AI Governance Framework for Agentic AI)[16]；澳洲偕同五眼聯盟發布《審慎導入代理式人工智慧服務》[17]，提醒最小權限、人為介入(human-in-the-loop)等共通原則。治理所關注的範圍，已從「它會不會答錯」擴大到「它能自己動手做多少事」。

代理本身也是可能被攻擊的弱點所在，政策因而必須跟上技術發展。對應的治理方向，是從代理的身分與授權著手：NIST 旗下國家網路安全卓越中心(NCCoE)的概念文件主張把既有身分協定與模型上下文協定(Model Context Protocol, MCP)納入代理治理，並將「可稽核性」列為必要條件[18]；德國聯邦資訊安全局(BSI)的查核表則援引「代理人雙重法則」(Agents Rule of Two)[19]，避免單一代理同時握有讀取機敏資料、接收不可信輸入與對外執行三種權限。最小權限與可稽核性，正成為公務機關導入代理時的政策底線。

代理與前沿模型的能力躍進，也開始牽動國家層級的因應。日本的「Project YATA-Shield」即是一例：以 Anthropic 前沿模型 Claude Mythos Preview 展現的自主漏洞挖掘能力為觸發點，日本分別針對關鍵基礎設施業者、軟體廠商與政府機關發布分眾警示與修補要求，並規劃由日本AI安全研究所與英國等外國同類機構協作評估前沿模型，形成把「前沿模型能力」直接連動到國家資安回應的機制[20]。韓國則向約2.8萬家企業資安長(CISO)發出強化要求，並提出以自主技術為基礎的「AI資安主權」目標[21]。近日，美國CISA偕G7多國發布的《AI軟體物料清單最低要素》(SBOM for AI)，則嘗試為代理與AI供應鏈的透明化建立共通的揭露格式[22]。

小結

綜觀上半年，各國做法雖有差異，其方向卻相近：在確保AI主權的前提下同步提升AI安全與資安，並讓治理由原則宣示走向制度落地、由規範模型輸出走向治理自主系統。我國其實也在同一波浪潮上：《人工智慧基本法》已於今年一月公布施行，因此，當前課題已不在「要不要立法」，而在如何落實，包括風險分類框架的建立、各機關的AI風險評估與使用規範等。

前述國際趨勢正可作為落地階段的參照：在量測的部分，各國多已投入資源建置AI安全研究機構與測試量能，值得我國及早布局；代理AI的身分、權限與可稽核性等，宜研究如何導入相關的安全規範；治理方向上，則需要持續思考如何強化資安事件監控與事後韌性等面向。

參考文獻

1. 國家資通安全研究院，《國際資安政策法制觀測》：https://www.nics.nat.gov.tw/core_business/information_security_information_sharing/International_Cybersecurity_Policy_Observation/
2. 美國白宮，《促進先進人工智慧創新與安全》行政命令(2026 年 6 月 2 日)：https://www.whitehouse.gov/presidential-actions/2026/06/promoting-advanced-artificial-intelligence-innovation-and-security/
3. Digital Omnibus on AI修法提案(歐盟理事會談判立場，2026 年 3 月 13 日)：https://www.consilium.europa.eu/en/press/press-releases/2026/03/13/council-agrees-position-to-streamline-rules-on-artificial-intelligence/
4. 歐盟執委會，《歐洲科技主權包裹》(2026 年 6 月 3 日)：https://ec.europa.eu/commission/presscorner/detail/en/ip_26_1187
5. 韓國《人工智慧基本法》(2026 年 1 月 22 日施行)：https://www.msit.go.kr/eng/bbs/view.do?bbsSeqNo=42&mId=4&mPid=2&nttSeqNo=1071&pageIndex=&sCode=eng&searchOpt=ALL&searchTxt=
6. 日本 J-AISI《Chief AI Officer 指引》(2026 年 3 月)：https://aisi.go.jp/assets/pdf/caio-guidebook.pdf
7. 韓國 MSIT，《人工智慧基本法支援窗口案例集》(2026 年 3 月 31 日)：https://www.msit.go.kr/bbs/view.do?sCode=user&mId=307&mPid=208&bbsSeqNo=94&nttSeqNo=3187094
8. 韓國 MSIT，人工智慧基本法施行令修正案(配合 7 月 21 日生效之修正法，2026 年 5 月 21 日預告)：https://www.msit.go.kr/bbs/view.do?sCode=user&mId=307&mPid=208&bbsSeqNo=94&nttSeqNo=3187340
9. 日本數位廳，《生成式 AI採購與利用指引》(DS-920)2.0 版(2026 年 6 月 12 日)：https://www.digital.go.jp/resources/standard_guidelines
10. UK AI Security Institute, Measuring AI Agents' Progress on Multi-Step Cyber Attack Scenarios(2026 年 3 月)：https://www.aisi.gov.uk/research/measuring-ai-agents-progress-on-multi-step-cyber-attack-scenarios
11. CAISI 等, Insights into AI Agent Security from a Large-Scale Red-Teaming Competition(arXiv:2603.15714，2026 年 3 月 23 日)：https://www.nist.gov/blogs/caisi-research-blog/insights-ai-agent-security-large-scale-red-teaming-competition ；CAISI 與 Google DeepMind、Microsoft、xAI之國家安全測試合作協議(2026 年 5 月 5 日)：https://www.nist.gov/news-events/news/2026/05/caisi-signs-agreements-regarding-frontier-ai-national-security-testing
12. JP-AISI、IPA，「JAI-Trust」大會(2026 年 5 月 21 日)：https://aisi.go.jp/activity/activity_information/260529/
13. 英國(AI Security Institute)與澳洲(Australian AI Safety Institute)AI安全研究院合作備忘錄(UK and Australia Pact on Fast-Moving AI Security Risks，2026 年 5 月 25 日)：https://www.gov.uk/government/news/uk-and-australia-pact-on-fast-moving-ai-security-risks
14. A. Vassilev, "Robust AI Security and Alignment: A Sisyphean Endeavor?", IEEE Security & Privacy(2026 年)；NIST 新聞稿(2026 年 6 月 9 日)：https://www.nist.gov/news-events/news/2026/06/nist-mathematical-proof-supports-transition-continuous-monitor-and-update
15. NIST, Challenges to the Monitoring of Deployed AI Systems(NIST AI800-4，2026 年 3 月)：https://www.nist.gov/news-events/news/2026/03/new-report-challenges-monitoring-deployed-ai-systems ；UK AI Security Institute, Loss of Oversight: How AI systems may become harder to audit, monitor, and investigate(2026 年 5 月)：https://www.aisi.gov.uk/research/loss-of-oversight-how-ai-systems-may-become-harder-to-audit-monitor-and-investigate
16. IMDA, Model AI Governance Framework for Agentic AI(2026 年 1 月 22 日)：https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/press-releases/2026/new-model-ai-governance-framework-for-agentic-ai
17. ASD / ACSC 偕五眼聯盟六機構，Careful Adoption of Agentic AI Services(2026 年 5 月 1 日)：https://www.cyber.gov.au/business-government/secure-design/artificial-intelligence/careful-adoption-of-agentic-ai-services
18. NIST NCCoE, Accelerating the Adoption of Software and AI Agent Identity and Authorization(概念文件，2026 年 2 月 5 日)：https://csrc.nist.gov/pubs/other/2026/02/05/accelerating-the-adoption-of-software-and-ai-agent/ipd
19. BSI, Evasion Attacks auf LLMs: Eine Checkliste zur Härtung des LLM-Systems(2025 年 11 月)：https://www.bsi.bund.de/DE/Service-Navi/Presse/Alle-Meldungen-News/Meldungen/Evasion-Attacks-LLM_251110.html ；「代理人雙重法則」見 Meta, Agents Rule of Two(2025 年 10 月 31 日)：https://ai.meta.com/blog/practical-ai-agent-security/
20. 日本內閣官房國家網路安全辦公室(NCO)聯合 14 省廳，Project YATA-Shield(2026 年 5 月 18 日)：https://www.cyber.go.jp/pdf/press/20260518_AI_CS_Package.pdf
21. 韓國 MSIT，第 9 次科學技術關係長官會議「AI基礎網路威脅因應民間資訊保護推進計畫」及「AI資安主權」目標(2026 年 5 月 29 日)：https://www.msit.go.kr/bbs/view.do?sCode=user&mId=307&mPid=208&bbsSeqNo=94&nttSeqNo=3187379
22. CISA、NCSC 及 G7 多國機構，Software Bill of Materials for Artificial Intelligence – Minimum Elements(2026 年 6 月)：https://www.cisa.gov/resources-tools/resources/software-bill-materials-ai-minimum-elements

關鍵字：國際AI治理、AI安全、代理型AI、AI法規政策

刊　　名　資安週報第 51 期
發 行 人　國家資通安全研究院　林盈達院長
主　　編　國家資通安全研究院　國際合作及資安治理中心
出 版 者　國家資通安全研究院
網　　址　www.nics.nat.gov.tw
訂閱網址　www.nics.nat.gov.tw/newsletter/
讀者信箱　www.nics.nat.gov.tw/mail2center/

國家資通安全研究院
National Institute of Cyber Security