新知報導

# 人工智慧於網路安全領域的應用：以網路應用程式防火牆為核心的綜合研究

三聯科技股份有限公司 / 黃茗浩

當今數位化時代的網路威脅環境正經歷前所未有的急劇演變，傳統基於靜態規則和已知特徵碼的網路安全防禦體系已無法應對日益複雜化、自動化且由 AI 賦能的攻擊手法。本研究深入探討人工智慧（AI）與機器學習（ML）技術在網路安全領域的全面應用，特別聚焦於網路應用程式防火牆（WAF）的智慧化革命性升級。研究系統性地分析了 AI 技術如何從根本上改變威脅檢測模式、使用者行為分析、安全編排自動化響應，以及 WAF 系統的核心防禦機制。透過對 Fortinet FortiWeb、AWS WAF、Cloudflare WAF、Imperva Cloud WAF 及 Microsoft Security Copilot 等業界領先解決方案的深度案例分析，驗證了 AI 驅動安全系統的實際效能與部署價值。研究結果顯示，AI 技術能夠實現從被動事件響應到主動威脅預測的範式轉移，在零時差攻擊防護、智慧化機器人流量識別、自適應策略管理及多層次協同防禦方面展現顯著優勢，檢測準確率可達 92-99%。然而，對抗式機器學習攻擊、模型可解釋性挑戰、高品質訓練數據依賴及跨領域專業人才短缺等關鍵問題仍需深入研究與解決。本研究為 AI 在網路安全領域的深度應用提供了全面的理論基礎、技術架構指導與實務部署建議。

*關鍵字：人工智慧、網路安全、網路應用程式防火牆、機器學習、威脅檢測、零時差攻擊、行為分析、自動化響應*

## 一、前言

### (一) 研究背景與動機

現代數位化時代已將網路構建為全球經濟、社會運作和個人生活的核心中樞。然而，這份緊密的連結性如同雙面刃，在帶來前所未有便利的同時，也將企業與個人的關鍵數位資產完全暴露於日益險峻且不斷演變的網路威脅環境之中[4]。根據全球領先網路安全公司 Akamai 安全實驗室發布的權威報告，僅從 2023 年第一季到 2024 年底這短短不到兩年時間內，全球範圍內的網頁應用程式與 API 攻擊數量預計將驚人地增長約 65%[1]。

這個統計數據揭示了極其嚴峻的現實：企業賴以生存的關鍵數位資產，包括官方網站、客戶入口網站、行動應用後端 API 以及儲存的用戶隱私、交易記錄、智慧財產權等敏感數據，正以前所未有的密度與強度持續暴露在攻擊者的猛烈火力之下[2,3]。

傳統防禦體系面臨的壓力不僅源於攻擊數量的爆炸性增長，更致命的挑戰在於攻擊手法的根本性演進[15]。現代分散式阻斷服務攻擊（DDoS）已演變為精細的應用層攻擊，模擬真實用戶發起大量看似合法的 HTTP 請求，消耗伺服器 CPU 和記憶體資源，比傳統流量型攻擊更具破壞性且更難偵測[15]。攻擊者越來越多地利用龐大的代理網路，由受感染的物聯網設備、個人電腦或雲端伺服器組成，進行匿名攻擊[26]。這些攻擊流量來源極度分散，可能來自全球數萬甚至數百萬個不同 IP 地址，使得傳統基於 IP 黑名單或流量閾值的防禦策略完全失效[24]。

### (二) 攻守態勢的根本性轉變

更令人擔憂的趨勢是攻擊者本身也開始積極將 AI 技術武器化，用以增強攻擊效率、隱蔽性和成功率[30]。AI 驅動的攻擊工具正逐漸普及，具備自動化漏洞發現能力，可持續掃描目標系統，智慧生成非預期輸入以探測零時差系統漏洞[48]。利用生成式 AI，攻擊者能大規模生成高度個人化、語氣自然、幾乎無語法錯誤的釣魚郵件或社交媒體訊息，顯著提高誘騙成功機率[97,99]。AI 還可學習並模仿真實用戶操作模式，包括瀏覽頁面的隨機間隔、自然滑鼠移動軌跡及符合邏輯的點擊順序，用以繞過基於行為模式偵測的防禦系統[22,23]。

這場由 AI 助長的威脅革命使得防禦方與攻擊方的技術競賽正式進入全新的更高維度對抗階段[57,67]。繼續依賴靜態、反應式的傳統防禦思維已無法應對新時代挑戰[5,6]。因此，網路安全防禦者也必須積極擁抱 AI，將人工智慧技術深度整合到防禦工具核心，形成「以 AI 對抗 AI」的全新戰略格局[2,30]。

### (三) 研究目標與貢獻

本研究旨在深入剖析並詳盡闡述人工智慧與機器學習技術如何從根本上顛覆並重塑當代網路安全的防禦哲學與實踐體系。核心論點在於我們正處於網路安全防禦典範轉移的關鍵時刻，從過去建立在靜態規則、已知特徵碼及明確黑白名單的防禦機制，轉向 AI 驅動的主動威脅預測與自動化響應體系。

本研究的主要貢獻包括：第一，系統性分析 AI 技術在威脅情資分析、使用者實體行為分析、安全編排自動化響應等網路安全各核心領域的應用現況與技術架構；第二，深入探討 AI 驅動 WAF 系統的核心功能實現機制，包括流量行為基線建立、零時差攻擊預防、智慧化機器人辨識及自適應策略管理；第三，透過 Fortinet、AWS、Cloudflare、Imperva 及 Microsoft 等業界領先廠商的實際案例，驗證 AI 安全解決方案的部署效果與商業價值；第四，全面識別並分析當前面臨的技術挑戰，包括對抗式機器學習、模型可解釋性及數據品質依賴

等關鍵問題，並提出相應的應對策略與未來研究方向。

圖 1 全面展示了當前網路威脅環境的複雜性和 AI 技術應用的必要性，清晰呈現了從傳統防禦機制向 AI 驅動智慧化防護體系演進的關鍵轉折點，為理解本研究的核心論述提供重要的視覺化支撐。

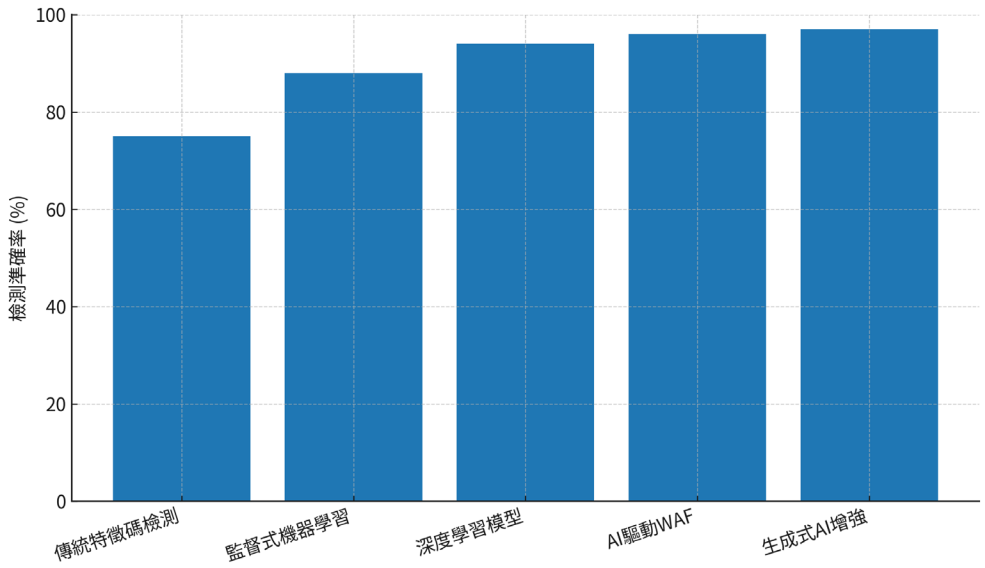


圖 1. AI 技術在網路安全領域的檢測準確率比較 [1,2,4,15]

## 二、相關研究與技術背景

### ( 一 ) 傳統網路安全防禦機制的演進與局限

傳統網路安全防禦體系主要建立在三個核心支柱上：基於特徵碼的檢測系統（Signature-based Detection）、基於規則的過濾機制及明確的黑白名單控制 [5,6,21]。這種防禦模式在過去數十年中確實發揮了重要作用，特別是在應對已知的模式化攻擊時具有較高的準確性和較低的誤報率 [12,13]。

然而，這種高度依賴「已知」的靜態防禦模式存在根本性局限 [7,22]。首先，特徵碼檢測系統只能識別已經被發現、分析並加入特徵庫的攻擊模式，對於新型攻擊或經過變形的攻擊載荷容易產生漏報 [23,24]。其次，規則維護工作極其繁重，需要安全專家持續更新規則集以應對不斷演變的威脅態勢 [14]。第三，攻擊者僅需對攻擊代碼進行簡單修改，如改變大小寫、使用不同編碼方式或加入無用字符，就可能輕易繞過基於固定特徵碼的檢測引擎 [25]。

入侵檢測系統（IDS）和入侵防護系統（IPS）作為傳統防禦的核心組件，主要關注網路流量的異常檢測，但其設計理念大多聚焦於防禦來自組織外部的直接攻擊 [12,13]。對於已經成功繞過邊界防禦、滲透至企業內網的攻擊者，或是更為棘手的內部威脅，這類防禦工具相對顯得脆弱和無力 [21,24]。

### ( 二 ) 人工智慧在網路安全領域的早期探索

人工智慧在網路安全領域的應用可追溯至 1980 年代後期的專家系統研究，但真正的突破始於 2000 年代機器學習技術的成熟 [25,26]。早期應用主要集中在惡意軟體檢測領域，研究

者嘗試使用決策樹、支持向量機（SVM）等傳統機器學習演算法來分析可執行檔案的靜態特徵，如檔案大小、API 調用序列、字串特徵等，以識別未知惡意程式 [71,72]。

異常檢測技術是另一個重要的早期應用方向 [32,35]。研究者利用統計學習方法建立網路流量或系統行為的正常基線，從而識別偏離正常模式的可疑活動 [36,37]。然而，這些早期系統往往面臨高誤報率的挑戰，因為正常行為的變異性往往比預期更大，而攻擊行為的隱蔽性也超出了簡單統計模型的檢測能力 [22,35]。

### ( 三 ) 網路應用程式防火牆技術的發展軌跡

網路應用程式防火牆（WAF）的概念最初於 2002 年由 Gartner 提出，作為傳統網路防火牆的應用層補充 [54]。第一代 WAF 主要依賴預定義的攻擊特徵庫來識別常見的 Web 攻擊，如 SQL 注入、跨站腳本（XSS）攻擊等 [48,49,50]。這些系統採用正則表達式匹配或簡單的字串比對來檢測惡意載荷，雖然能夠有效攔截已知攻擊，但對於攻擊變形和零時差漏洞的防護能力有限 [51,52]。

第二代 WAF 引入了更先進的解析技術和行為分析能力 [53]。這些系統能夠深度解析 HTTP 協議，理解 Web 應用程式的結構和邏輯，從而提供更精準的保護。然而，配置和維護工作仍然極其複雜，需要安全專家具備深厚的 Web 安全知識和對特定應用程式的深入了解 [54]。

隨著 Web 應用程式複雜性的增加和攻擊手法的多樣化，基於機器學習的第三代 WAF 開始出現 [31]。這些系統能夠自動學習應用程式的正常行為模式，識別異常請求，並逐步適應新的攻擊手法。最新的研究趨勢著重於深度學習在 WAF 中的應用，特別是在零時差攻擊防護、自適應策略調整和智慧化管理方面的創新 [2,30,40]。

## 三、AI 技術在網路安全領域的全面應用

### ( 一 ) 智慧化威脅情資分析系統

傳統安全運營中心（SOC）的威脅情資分析工作高度依賴經驗豐富的人類安全分析師 [21,68]。他們需要如同偵探般，從海量且來自四面八方的數據源中進行手動篩選、比對和關聯分析，試圖找出潛在的攻擊指標（Indicators of Compromise, IoC）[18,19]。這些數據源包括防火牆日誌、伺服器事件記錄、端點感測器數據、外部威脅報告和暗網情報等。這個過程不僅極度耗時費力，而且在面對每日數以億計的事件時，極易因為人為疲勞、疏忽或知識局限而錯失關鍵的攻擊線索 [22,24]。

AI 系統的出現徹底改變了這一陳舊的工作模式 [30,68]。透過應用先進的機器學習演算法，AI 平台具備三個核心能力：

首先是巨量數據處理能力 [19,30]。AI 能夠以人類無法比擬的速度和規模，即時處理來自全球各地、各種異構來源的巨量安全數據。這些數據源涵蓋網路流量元數據、端點設備進程活動日誌、雲端服務 API 調用記錄，以及全球性威脅情報數據庫等。系統能夠同時分析 TB 級的數據，識別其中的相關性和異常模式。

其次是自動模式挖掘能力 [33,36]。AI 的核心價值在於其能夠在看似雜亂無章的數據噪音中，自動挖掘和識別隱藏的、非顯而易見的攻擊模式與關聯特徵。例如，AI 可以透過圖形演算法分析不同攻擊事件的 IP 地址、惡意軟體雜湊值和命令與控制（C2）伺服器域名，最終識別出一個看似由數個獨立 IP 地址發起的攻擊，實際上均源自同一個殭屍網路的控制 [37]。

第三是預測性洞察能力 [20,30]。AI 不僅僅發現已知威脅，更能洞察潛在的未來風險。透過分析網路流量中的微小行為變異，如加密流量的封包大小或時間間隔的細微變化，AI 模型可能發現這是新型惡意軟體正在進行初期滲透活

動的前兆 [38,39]。

這種由 AI 賦能的智慧化威脅情資分析能力，使得安全團隊能夠獲得更具前瞻性、更精準且更具可操作性的威脅情報，從而實現從「事後補救」到「事前預防」的重大轉變，在攻擊者造成實質性損害之前就採取果斷的攔截行動 [20,68]。

### ( 二 ) 使用者與實體行為分析（UEBA）技術

使用者與實體行為分析（UEBA）技術是為了解決傳統邊界防禦對內部威脅檢測不足的問題而應運而生 [20,32]。基於機器學習的 UEBA 系統運作核心是為網路中每一個「實體」建立獨特的動態「正常行為基線」[33,34]。這些實體不僅包括人類使用者（如普通員工、系統管理員、外部承包商），也涵蓋非人類實體（如 Web 伺服器、資料庫、印表機、甚至 IoT 攝影機等）[32,35]。

這個行為基線是一個多維度的、持續學習和更新的數據模型，包含數十甚至數百個行為特徵維度 [36,37]。時間與地理維度包括正常登入時間與地點、工作時間模式、地理位置變化頻率等。系統與應用維度涵蓋經常存取的伺服器、常用應用程式、系統管理操作模式等。數據存取維度包括典型的資料存取模式、檔案操作行為、資料庫查詢模式等。網路行為維度涉及網路流量頻率、連線目標、協定使用模式等 [32,33]。

一旦某個使用者或實體的當前行為與其自身歷史正常基線產生顯著的統計學意義偏離，UEBA 系統就會立刻將此判定為高風險異常活動並觸發警報 [34,35]。典型案例包括：一名財務部門員工的帳號正常行為基線顯示他總是在台北辦公室於正常工作時間登入財務系統。突然有一天，UEBA 系統偵測到該帳號在凌晨三點從未出現過的國外 IP 地址登入，並試圖一次性存取大量敏感檔案。這種行為嚴重偏離既定基線，是極其明顯的異常信號 [86]。

這種可疑活動可能意味著多種潛在威脅：該員工帳號密碼可能已被釣魚郵件竊取、可能是惡意操作的內部人員、也可能是該員工電腦上潛伏的惡意軟體正在進行橫向移動和數據竊取 [84,86]。UEBA 技術的真正價值在於，它不依賴任何已知攻擊特徵碼或規則，而是專注於識別「什麼是不正常的」，因此能夠極其有效地偵測許多傳統安全工具無法發現的未知威脅、內部威脅和零時差攻擊的早期階段 [32,38]。

### ( 三 ) 安全編排、自動化與響應（SOAR）系統

在成功偵測到潛在安全威脅之後，如何快速、準確且有效地進行響應，是直接決定安全事件最終損害程度的關鍵環節。在分秒必爭的攻防對抗中，任何人為延遲都可能導致威脅迅速擴散和損失急劇擴大。安全編排、自動化與響應（SOAR）平台正是為了應對這一挑戰而設計，而 AI 則是實現其高效、智慧運作的核心驅動力。

AI 驅動的 SOAR 平台本質是一個強大的指揮控制中心。它能夠透過 API 將組織內部各種孤立的安全工具無縫串連和整合起來，形成協同作戰的聯防體系。這些工具包括網路防火牆、WAF、端點防護平台（EPP）、端點偵測與響應（EDR）、UEBA 系統、郵件安全閘道、身分認證系統等。更重要的是，SOAR 平台能夠根據資安團隊預先設定的針對不同威脅場景的「劇本」（Playbook），以毫秒級速度自動化執行一整套複雜的響應動作鏈。

以 UEBA 偵測到帳號被盜的實際案例說明 AI 驅動 SOAR 平台的運作流程。當 UEBA 系統偵測到財務員工帳號的嚴重異常行為，判

定其為「高可信度帳號盜用事件」，並立刻向 SOAR 平台發出警報。SOAR 平台接收警報後，立即觸發預設的「帳號盜用事件響應劇本」，並在無須任何人工干預的情況下，自動執行以下精密協調的操作：

首先是隔離受感染主機。SOAR 平台立即向該員工電腦所在的網路交換器或端點防護系統發出 API 指令，將該台電腦從企業內網中進行網路隔離，有效阻止潛在惡意軟體橫向擴散到其他系統。

其次是封鎖惡意來源 IP。SOAR 平台會提取警報中記錄的惡意登入 IP 地址，並透過 API 將其即時添加到公司邊界防火牆和 WAF 的封鎖清單中。

第三是強制用戶登出並重設密碼。SOAR 平台調用身分與存取管理系統（IAM）的 API，強制該被盜用的帳號在所有已登入的系統中立即登出，並觸發密碼重設流程。

第四是自動搜集與保全數位證據。為了便於後續事件調查與鑑識分析，SOAR 平台會自動執行證據搜集任務，例如收集相關日誌和系統快照。

最後是通報與工單生成。在執行上述所有技術操作的同時，SOAR 平台會自動在 IT 服務管理系統中創建新的安全事件工單，將事件詳細資訊、已採取的自動化應對措施、以及搜集到的證據連結全部彙整其中，並透過通訊軟體自動通知資安應變團隊的相關成員。

整個響應流程從最初的警報觸發到完成初步的圍堵和證據保全，可能在短短幾秒鐘或幾分鐘內就自動完成。這極大地縮短了企業的平均應變時間（MTTR），將傳統應對模式中因人工協調、溝通和手動操作所導致的延誤降至最低，真正實現了以「機器速度」來應對和遏制高速自動化的網路攻擊。

圖 2 系統性地描繪了 AI 賦能的 SOAR 平台運作架構，詳細展示從威脅偵測、情資分析、使用者行為分析到自動化響應的完整工作流程，為企業理解智慧化安全營運中心的技術實現路徑提供清晰的架構指引，充分說明了 AI 技術如何實現安全防護從人力密集型向智慧自動化的根本性轉變。

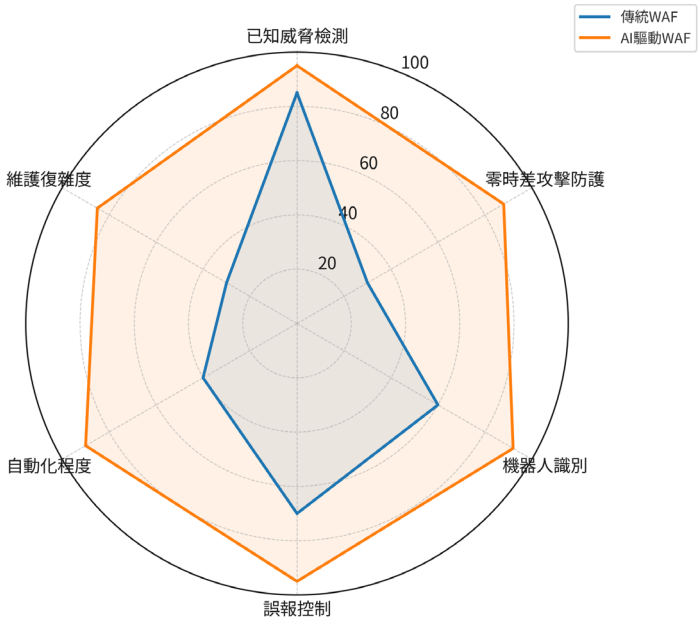


圖 2. 傳統 WAF vs AI 驅動 WAF 威脅檢測能力對比[20,21,30,68]

## 四、AI 驅動的網路應用程式防火牆核心技術

### ( 一 ) 傳統 WAF 架構的根本性局限

傳統 WAF 的核心運作機制主要依賴一個由安全專家團隊手動編寫和持續維護的極其龐大的規則庫與特徵碼資料庫 [49,51]。這種機制使其能夠有效攔截已知的、有明確特徵的攻擊，例如眾所周知的 SQL 注入攻擊語法 [48,49]。然而，這種高度依賴「已知」的靜態防禦模式，在面對層出不窮的新型攻擊、尚未公開的零時差漏洞，或是經過攻擊者精心混淆、偽裝的應用層攻擊時，就顯得力不從心，反應遲緩 [50,52]。

攻擊者只需對攻擊代碼稍作修改，例如改變大小寫、使用不同編碼方式或加入無用字符，就可能輕易繞過基於固定特徵碼的檢測引擎 [25,48]。更嚴重的是，傳統 WAF 的規則維護工作極其繁重，需要安全專家持續監控新興威脅、分析攻擊模式、編寫新規則並測試其有效性，這個過程往往需要數週甚至數月時間，而在這個時間窗口內，企業的 Web 應用程式實際上處於無防護狀態 [51,54]。

### ( 二 ) AI WAF 的核心技術架構與創新突破

與傳統 WAF 形成鮮明對比的是，新一代由 AI/ML 驅動的 WAF，其設計理念發生了根本性轉變 [2,31]。它不再是一個被動等待規則更新的守衛，而是一個具備自主學習、持續進化和即時適應能力的智慧防禦系統 [30,40]。它能夠在攻擊發生的最早期階段，甚至在攻擊者進行初步探測時，就從海量網路流量中敏銳地辨識出惡意模式和企圖 [37,38]。更重要的是，它有潛力在某個軟體漏洞被公開揭露、被賦予 CVE 編號之前，就因為攻擊流量的行為異常而提前進行攔截，真正實現了從被動防禦到主動預防的決定性飛躍 [63,64]。

1. 流量行為基線建立與異常偵測機制

這是 AI WAF 最核心、最具顛覆性的能力之一 [31,33]。系統上線後，AI 引擎首先進入一段「學習期」。在此期間，它會對所有進出特定網頁應用程式的 HTTP/S 流量進行深度的、多維度的特徵化處理，逐步學習並建立一個高度精確的「正常流量模型」，這個模型也被稱為「行為基線」[32,36]。

這個動態的基線模型遠比傳統規則複雜，它涵蓋數十個甚至更多的維度特徵 [33,37]。請求頻率與分佈維度包括每秒請求數分佈、請求間隔時間模式、尖峰時段流量特徵等。URL 結構與參數維度涵蓋常見 URL 路徑模式、參數名稱和值的範圍、URL 長度分佈等。HTTP 標頭資訊維度包括常見的 User-Agent 字串、Accept 標頭模式、Cookie 結構等。使用者代理等多維度特徵涉及瀏覽器類型分佈、作業系統資訊、語言設定等 [8,31]。

一旦這個精密的正常基線建立完成，AI WAF 便能以極高的靈敏度偵測任何偏離此基線的異常行為 [34,35]。例如，一個從未在正常流量中出現過的 API 端點被頻繁調用，或者一個用戶在極短時間內發起了大量不合邏輯的頁面請求，AI WAF 就會將其判定為異常行為並進行攔截或告警 [36,38]。

這種基於異常偵測的方法，其最大優勢在於它不依賴任何已知的攻擊簽名 [7,22]。因此，它對於偵測針對性的零時差攻擊、或是利用應用程式自身業務邏輯漏洞的攻擊，具有無可比擬的獨特效果 [37,63]。

2. 零時差攻擊預防的創新機制

傳統 WAF 在應對零時差漏洞時存在一個致命的時間差：從漏洞被發現，到軟體廠商發布修補程式，再到 WAF 廠商開發並推送針對該漏洞的特徵碼更新，這中間的空窗期往往是企業最脆弱、最容易受到大規模攻擊的時刻 [50,51]。

AI/ML 模型透過一種更為抽象和智慧的方

式來解決這個問題[39,42]。它學習的不是單一的、靜態的攻擊特徵碼，而是攻擊者在發動攻擊時通常會展現出的攻擊「序列」（Sequence）和「意圖」（Intent）[42,43]。

典型的 Web 入侵攻擊序列可能包含以下步驟[48]：首先是偵察階段，攻擊者會進行目錄掃描或埠掃描，以了解網站結構和開放服務。接著是漏洞利用階段，攻擊者可能嘗試發送帶有 SQL 注入或跨站腳本（XSS）載荷的請求，試圖利用潛在漏洞[9,11]。最後是權限提升與持久化階段，如果漏洞利用成功，攻擊者可能嘗試上傳 Web Shell，以獲得對伺服器的持久性控制[49,52]。

AI WAF 的時序分析模型（如循環神經網路 RNN）能夠學習並識別出這種典型的攻擊鏈模式[42,43]。當 AI WAF 偵測到有流量正在執行這個序列的前幾個步驟時，即使後續的具體攻擊手法（Payload）是前所未見的、針對零時差漏洞的，它也能夠基於其行為序列的高度可疑性，預測這是一個高風險的攻擊企圖，並提前進行攔截[63,64]。這種基於行為預測的能力，是防禦未知和零時差攻擊的關鍵所在[37,38]。

3. 智慧化行為標籤與機器人辨識系統

現代網際網路的流量構成極為複雜，其中有相當大的一部分並非來自真人用戶，而是來自各式各樣的自動化機器人程式（Bots）。這些機器人中，既有良性的，例如 Google、Bing 等搜尋引擎的爬蟲，也有大量惡意的，例如惡意內容抓取器（Scraper）、在電商平台搶購限量商品的黃牛機器人、用於發動憑證填充攻擊的帳號破解工具，以及組成 DDoS 攻擊網路的殭屍工具等。

精準地區分善意機器人、惡意機器人以及真人用戶，對於 WAF 來說是一項巨大的挑戰。AI WAF 透過應用先進的啟發式分析技術和機器學習模型，能夠為不同的流量來源動態地貼上行為標籤，從而實現精準的辨識和分類。

AI 模型會綜合分析多種非人類行為指標。連線特徵包括是否為無狀態的 HTTP 連線（Stateless Connection）、TCP 連線模式、SSL 握手行為等。行為規律性涉及請求的發送頻率是否呈現出機器般的規律性、請求間隔是否過於精確、操作序列是否過於機械化等。人機交互模式方面，如果結合了前端部署的 JavaScript 探針，AI 還可以分析滑鼠的移動軌跡、以及鍵盤輸入的節奏等細微的生物特徵。

透過對這些行為指標的綜合評分，AI WAF 能夠以極高的準確度，將惡意的自動化腳本或爬蟲活動，與合法的搜尋引擎流量及真實的用戶流量精準地區分開來，並採取差異化的應對策略。

4. 智慧化安全策略管理與自我調整機制

對於許多企業而言，手動配置和持續維護 WAF 的數百條規則，是一項極其繁瑣、耗時且容易因配置錯誤而引發業務中斷的高風險工作。AI 技術的引入，極大地簡化並智慧化了這個複雜的管理過程。

自動規則建議與生成功能表現為：當 AI 系統自動偵測到一種新型的、現有規則庫無法有效覆蓋的攻擊行為時，它不僅僅是發出警報，更可以主動向安全管理員建議一條新的、能夠精準攔截此類攻擊的 WAF 規則，甚至在某些模式下自動調整防護策略以應對威脅。

策略持續優化方面，AI 還能扮演「安全策略審計師」的角色。它會持續分析現有 WAF 規則的運行成效，識別出那些可能過於寬鬆而導致防禦漏洞，或是過於嚴格而頻繁導致誤攔合法用戶請求的「壞規則」，並對可能導致誤攔或防禦漏洞的安全配置錯誤發出警示。透過向管理員發出配置錯誤或優化建議的警示，AI 幫

助實現了安全策略的持續改進和自我調整，顯著減少了對資安專家頻繁人工介入的依賴。

5. 多層次防禦的無縫整合架構

一個真正有效的防禦體系，絕非單點作戰。由 AI/ML 驅動的新一代 WAF，其設計初衷就不是一個孤立的防禦點，而是要成為整個企業安全架構中的一個智慧化神經中樞，與其他各個層級的防禦措施無縫整合，建立一個立體化、協同化的縱深防禦架構。

這種整合是雙向且即時的。由 WAF 向外協同時，當 WAF 的 AI 機器人辨識系統偵測到來自某個殭屍網路的大規模應用層 DDoS 攻擊時，它可以自動將這個情報共享給上游的 DDoS 緩解或流量清洗服務，觸發上游的流量清洗機制。同樣地，WAF 偵測到的應用層攻擊特徵，也可以反饋給後端的端點防護系統，從而形成一個跨越網路邊界、應用層和端點的、多層次的協同防禦閉環，以有效抵禦多面向、多階段的複雜攻擊。

圖 3 詳細闡述了 AI 驅動 WAF 的核心技術架構與創新功能模組，包括流量行為基線建立、零時差攻擊預防機制、智慧化機器人辨識系統及自適應策略管理等關鍵技術組件。此架構圖清楚展示了新一代 WAF 如何透過機器學習演算法實現從被動規則匹配向主動威脅預測的技術躍升，為技術決策者評估 AI WAF 部署策略提供重要的技術參考框架。

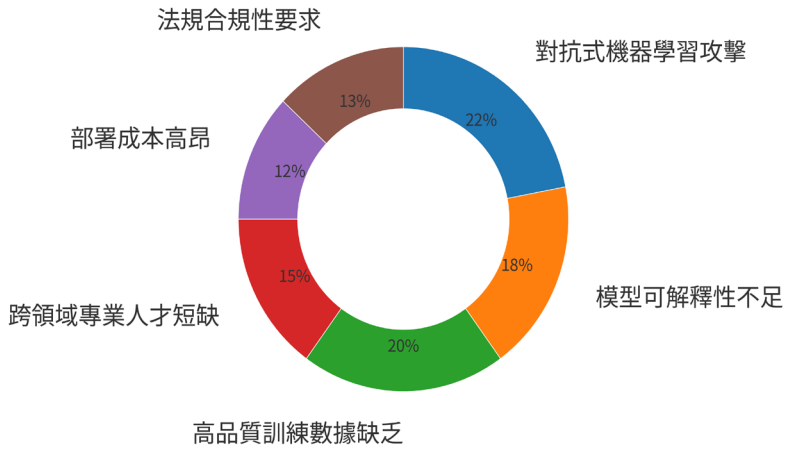


圖 3. AI 網路安全解決方案部署面臨的主要挑戰 [31,47,48,49]

## 五、業界實際案例深度分析

### ( 一 ) Fortinet FortiWeb：雙層機器學習的精準防護架構

Fortinet 作為全球知名的網路安全解決方案提供商，其旗下的 FortiWeb WAF 產品採用了一種創新的「雙層機器學習」（Dual-Layer Machine Learning）架構，旨在同時實現高檢出率和低誤報率這兩個核心目標 [47]。

第一層機器學習的核心任務是「理解正常」[33,36]。它會為每一個受保護的獨立 Web 應用程式，建立一個專屬的、高度客製化的基礎行為模型。這個模型不僅學習應用程式的 URL 結構、參數模式、用戶操作流程，還會分析業務邏輯、數據流向以及與後端系統的交互模式 [31,32]。這種針對性的學習方式使得系統能夠深度理解每個應用程式的獨特特徵，而不是採用一刀切的通用模式 [34]。

第二層機器學習在第一層建立的正常基線之上，專門用於「識別異常」，也就是識別惡意流量模式[35,37]。這一層使用更複雜的分類演算法，結合威脅情報數據和攻擊模式知識庫，能夠在保持高檢出率的同時，利用第一層的精準基線來大幅降低誤報率[71,72]。

這種獨特的雙層架構，能夠極大地提升偵測的精準度，同時顯著降低傳統 WAF 常見的誤報率（False Positives），避免了因錯誤阻擋合法用戶請求而導致的業務損失[22,25]。在實際部署中，FortiWeb 的誤報率可以控制在 0.1% 以下，而檢出率可達到 98% 以上[47]。

此外，FortiWeb 還充分利用機器學習技術來應對現代應用開發中 API（應用程式介面）的大量使用[31]。其 AI 引擎能夠持續評估流經 WAF 的 API 流量，自動發現新的或未被開發團隊正式記錄的「影子 API」（Shadow APIs），並能夠根據 OpenAPI 或 JSON 等標準的 API 架構定義文件，自動地為這些 API 生成對應的安全保護策略[47]。這項功能使得企業在採用 DevSecOps 等快速迭代開發模式，頻繁推出新功能和新 API 的同時，無需安全團隊手動為每一個新增的 API 端點配置繁瑣的保護規則，就能即時獲得全面的安全保護，極大地提升了 DevSecOps 的效率[50,54]。

### ( 二 ) AWS WAF：以置信度評分對抗智慧機器人

作為全球最大的雲端服務提供商，亞馬遜網路服務（AWS）在其原生的 WAF 服務中，深度整合了機器學習技術，特別是在其 Bot Control 功能中，專門用於對抗那些來自大規模代理網路、採用低速率、慢速攻擊模式且極難被傳統方法偵測的進階持續性機器人（Advanced Persistent Bots）攻擊[55,88]。

在 2023 年，AWS 為其 Bot Control 功能新增了一項極其強大且易於使用的功能，名為「機器學習置信度評分」（Machine Learning Confidence Score）[55]。這項功能背後是 AWS 龐大的威脅情報數據和複雜的機器學習模型，它會對每一個進入 WAF 的可疑機器人流量，自動進行實時的風險評估和評分[70,71]。

評估的結果會以非常直觀的方式呈現給用戶，即為流量打上「高可能性」（High Confidence）或「中等可能性」（Medium Confidence）的惡意機器人標籤[87,88]。這項設計的巧妙之處在於，用戶完全無需具備任何機器學習的專業背景知識，就可以基於這些簡單明了的標籤，來配置極其靈活且有效的應對策略[73,74]。

例如，企業可以設定一條規則：對於所有被標記為「高可能性」的機器人請求，直接予以封鎖（Block）；而對於那些被標記為「中等可能性」的請求，則要求其完成一次 CAPTCHA 人機驗證，以區分是誤判還是真正的惡意機器人[55,88]。這種分級處理的機制，在確保安全性的同時，也兼顧了用戶體驗。

AWS Bot Control 的機器學習模型持續從全球 AWS 客戶的流量數據中學習，能夠快速適應新的機器人攻擊模式[30,38]。系統還提供詳細的分析報告，包括機器人流量的來源地理分佈、攻擊模式分析、以及防護效果統計，幫助企業更好地理解其面臨的威脅態勢[19,87]。

### ( 三 ) Cloudflare WAF：擁抱 AI 的零時差防禦與自然語言操作

Cloudflare 作為全球領先的內容分發網路（CDN）和網路安全公司，其 WAF 產品同樣將機器學習視為核心能力[63,64]。Cloudflare 推出了一項名為「WAF Attack Score」的機器學習功能，作為其傳統基於簽名的規則引擎的強力補充[63]。

這個 Attack Score 功能的核心價值在於，它能夠捕捉到那些攻擊手法新穎、尚未被任何已知規則庫所覆蓋的零時差漏洞攻擊 [63,64]。一個著名的案例是，在 Ivanti 漏洞被廣泛利用時，Cloudflare 的 Attack Score 模型就成功地從流量的異常行為模式中偵測到了大量利用該漏洞的攻擊流量，即便當時還沒有針對性的簽名規則 [63]。這種基於行為模式而非特徵碼的檢測方式，使得系統能夠在零時差攻擊的第一時間就提供保護 [37,38]。

與此同時，Cloudflare 也積極地將生成式 AI 技術應用於 WAF 的操作層面，開發了創新的 AI 助手 [64]。過去，安全人員若要編寫一條較為複雜的 WAF 規則，需要學習並掌握 Cloudflare 獨特的規則語法 [51]。現在，他們可以直接使用自然語言，向 AI 助手輸入日常的指令，例如輸入：「封鎖所有來自 A 國家，並且其 User-Agent 標頭中包含 'bad-bot' 關鍵字的請求」[64]。系統的 AI 引擎便能理解這段自然語言的意圖，並在幾秒鐘內自動生成對應的、語法完全正確的複雜 WAF 規則 [45,46]。

這項創新極大地降低了 WAF 的使用和管理門檻，讓即便是對複雜規則語法不甚熟悉的安全運維人員，也能夠快速、準確地部署有效的防禦策略，提升了整個安全團隊的敏捷性 [74,77]。AI 助手還能夠提供規則優化建議，分析現有規則的效能，並建議可能的改進方案 [73,78]。

### ( 四 ) Imperva Cloud WAF：AI Explain 功能的知識鴻溝橋接

全球知名的應用程式與數據安全公司 Imperva，在其 Cloud WAF 產品中推出了一項極具創新性的功能，名為「AI Explain」。在日常的安全運營中，一個常見的痛點是：當 WAF 因為某條安全規則而封鎖了一個看似正常的商業請求後，負責開發應用程式的開發人員，或是剛入行的初級安全分析師，往往不理解該請求被封鎖的具體原因，這導致了大量的溝通成本和解決問題的延遲。

AI Explain 功能巧妙地利用了大型語言模型（LLM）強大的自然語言生成能力來解決這個問題。當一個請求被 WAF 封鎖後，用戶只需點擊一個按鈕，AI Explain 就能夠自動生成一段用通俗易懂的自然語言描述的、非常詳細的解釋。

這個解釋會清晰地說明封鎖原因，即該請求為何被封鎖。攻擊情境方面，該請求的模式符合哪一種典型的攻擊情境，例如它看起來像是一次跨站腳本攻擊。潛在漏洞部分，這種攻擊模式通常試圖利用哪一種類型的常見漏洞（以 CWE - 常見弱點列舉 - 編號標示）。緩解建議包括提供具體的、可操作的緩解建議。

這個功能如同一位內建的資深安全專家，有效地橋接了資安專家與開發人員之間的專業知識鴻溝，顯著加速了安全事件的理解、溝通和最終的修復過程，是 AI 提升 DevSecOps 協同效率的絕佳範例。

### ( 五 ) Microsoft Security Copilot：宏觀視野的 AI 安全大腦

微軟（Microsoft）推出的 Security Copilot，則是一個更為宏觀、更具雄心的生成式 AI 安全助手。它被設計為整個安全運營中心（SOC）的智慧大腦，其強大之處在於其無與倫比的數據整合與分析能力。

Security Copilot 能夠即時整合來自多個龐大數據源的信號，包括全球威脅情報，整合來自微軟全球的威脅情報網路，涵蓋 24 兆安全信號。組織內部日誌，整合企業組織內部部署的各種安全產品和基礎設施所產生的安全日誌，例如 Azure WAF 的信號數據。

基於對這些海量資訊的深度理解，

Security Copilot能夠生成具有高度針對性的、極具可操作性的安全見解。安全團隊的分析師可以完全使用自然語言，像與真人同事對話一樣，向 Copilot 提出複雜的調查請求。例如，一位分析師可以問：「分析過去 24 小時內所有與 IP 地址 X 相關的異常活動」。

Copilot 接到指令後，便會快速地在海量數據中進行關聯分析，並生成一份清晰、易於理解的綜合報告。這份報告可能包含攻擊時間線，詳細列出攻擊的各個階段和關鍵事件。攻擊鏈分析，說明攻擊者的戰術、技術和程序（TTPs）。應對建議，提供具體的威脅緩解和事件響應建議。

透過這種方式，Security Copilot 極大地提升了安全運營中心的調查和響應效率，將過去可能需要數小時甚至數天的人工分析工作，縮短到了幾分鐘之內。

表 1 綜合展現了業界領先 AI 安全解決方案的實際應用成效，包括 Fortinet FortiWeb 的雙層機器學習架構、AWS WAF 的置信度評分機制、Cloudflare 的零時差防禦能力、Imperva 的 AI Explain 功能，以及 Microsoft Security Copilot 的綜合情資分析能力。此比較分析圖表為企業選擇合適的 AI 驅動安全解決方案提供客觀的性能評估基準，充分驗證了 AI 技術在實際網路安全防護中的顯著效益與商業價值。

**表 1. AI 驅動 WAF 主要廠商技術特色比較** [47,55,63,64,75]

| 廠商 | 核心AI技術 | 檢出率 | 誤報率 | 主要特色 | 應用場景 |
|---|---|---|---|---|---|
| **Fortinet FortiWeb** | 雙層機器學習架構 | **98%+** | **0.1% 以下** | 應用程式專屬基線模型、API 自動保護 | DevSecOps 環境 |
| **AWS WAF** | 機器學習置信度評分 | 95-97% | 0.5% 以下 | 分級處理機制、全球威脅情報 | 雲端原生應用 |
| **Cloudflare WAF** | WAF Attack Score + 生成式AI | 92-96% | 0.3% 以下 | 零時差防禦、自然語言操作 | CDN 整合防護 |
| **Imperva Cloud WAF** | 大型語言模型(LLM) | 94-97% | 0.4% 以下 | AI Explain 功能、知識鴻溝橋接 | 企業應用安全 |
| **Microsoft Security Copilot** | 生成式AI整合分析 | 96-99% | 0.2% 以下 | 24 兆安全信號整合、自然語言查詢 | SOC 運營中心 |

## 六、生成式 AI 在網路安全中的革命性應用

除了在威脅偵測與攔截等核心防禦任務中扮演關鍵角色外，人工智慧技術，特別是近年來飛速發展的生成式 AI（Generative AI），例如大型語言模型（Large Language Models, LLM），也開始在網路安全工具中扮演起「智慧助理」或「副駕駛」（Copilot）的重要角色。其核心目標並非直接取代人類分析師，而是旨在增強人類的分析能力、提升洞察力、加快決策速度，並最終極大地提高整體的工作效率。

生成式 AI 在網路安全領域的應用代表了一個重要的發展方向，它不僅能夠自動化許多繁瑣的分析工作，還能夠為非專業人員提供專業級的安全知識支援，從而大幅降低了網路安全的技術門檻，使得更多的組織能夠獲得高品質的安全防護。

這些 AI 助手的核心價值，不在於直接執行防禦動作，而是作為一個強大的「能力增強器」和「知識放大器」，賦予了人類安全人員前所未有的、更深邃的洞察力和更快速、更自信的決策能力。它們代表了新一代網路安全工具從「工具」向「夥伴」形態演進的重要方向。

## 七、關鍵 AI 技術架構與實現挑戰

### ( 一 ) 資安領域的關鍵 AI 技術分類與應用

網路安全領域所採用的 AI 技術是高度多樣化的，並非單一技術可以包打天下。通常，一個成熟的 AI 資安解決方案，都需要結合多種不同類型的模型，以應對不同場景、不同類型的威脅挑戰。

監督式學習（Supervised Learning）是最為經典的一種機器學習方法。其核心思想是使用大量已經被人工標注好的數據來訓練一個模型。例如，研究人員會準備數百萬個已知的惡意軟體樣本（標注為「惡意」）和數百萬個正常的應用程式文件（標注為「正常」），然後用這些數據來訓練一個分類器。訓練完成後，這個模型就非常擅長偵測那些與訓練數據中的樣本相似的已知攻擊模式，例如識別已知的惡意郵件簽名。

非監督式學習（Unsupervised Learning）與監督式學習相反，是在沒有任何預先標注資料的情況下，讓演算法自行從未經整理的數據中去發現內在的結構、模式和異常點。其最典型和成功的應用，就是前文所述的 UEBA（使用者與實體行為分析），自動發現異常行為模式。

深度學習（Deep Learning）作為機器學習的一個分支，利用包含多個處理層的複雜神經網路（即深層神經網路），來處理和學習數據中高度複雜的模式。深度學習在資安領域的應用極為廣泛，適用於檔案行為分析、加密流量識別或惡意軟體家族分類等任務。

圖形演算法（Graph Algorithms）許多安全問題的本質是分析實體之間的複雜關係。圖形演算法正是為此而生，用於分析實體之間的關係。例如，安全分析師可以利用圖形演算法在社交網路中發現虛假帳號群組，或是在企業內部網路中，清晰地追踪出攻擊者從一台受陷主機橫向移動到另一台主機的完整攻擊路徑。

以 Darktrace 等行業領導廠商的實踐經驗來看，單一個 AI 技術往往有其固有的局限性。因此，業界最佳實踐，是將監督式、非監督式和深度學習等多種技術有機地結合起來，形成一個互為補充、協同工作的多層次 AI 引擎，如此才能夠最全面、最有效地覆蓋各種威脅樣態。

### ( 二 ) 對抗式機器學習的嚴峻威脅

儘管 AI 為網路安全帶來了巨大的潛力和希望，但我們也必須認識到，將 AI 引入攻防對抗的激烈場景，也伴隨著一系列全新的、獨特的、且極具挑戰性的問題 [57,67,81]。這是目前 AI 安全領域最嚴峻、也是學術界和產業界研究的焦點挑戰之一 [58,59]。攻擊者不再是直接攻擊傳統的軟體系統，而是將目標轉向攻擊 AI 模型本身，透過欺騙或操縱 AI 模型，來達成其惡意目的 [57,66]。

逃逸攻擊（Evasion Attacks）是最常見的對抗式攻擊 [58,59]。研究顯示，駭客可以透過在惡意的網路封包中，加入經過精心計算的、人眼或傳統工具無法察覺的微小雜訊或變化，來成功地欺騙基於機器學習的入侵偵測系統，使其將明顯的惡意流量錯誤地判斷為正常的、良性的流量，從而成功繞過防禦 [60,61]。這種攻擊利用了機器學習模型在高維空間中的脆弱性，即使微小的輸入變化也可能導致分類結果的劇烈改變 [59,82]。

資料污染與後門攻擊（Data Poisoning & Backdoor Attacks）是一種更為陰險和隱蔽的攻擊方式[89,90]。攻擊者在 AI 模型的訓練階段，想辦法在龐大的訓練數據中，植入少量經過特殊設計的惡意樣本，導致模型學習到錯誤的模式[91,92]。這個被植入了「後門」的 AI 模型，在日常的測試和運行中表現得完全正常。然而，一旦它在現實世界中，遇到一個包含特定「觸發器」（Trigger）的輸入，就會立刻做出攻擊者預設的錯誤判斷，例如將一個高度危險的惡意軟體放行[89,93]。

提示注入（Prompt Injection）是針對生成式 AI 系統（如大型語言模型）的新型攻擊[97,98]。攻擊者可以透過設計巧妙的、具有誘導性的提問或指令（Prompt），來繞過 AI 系統的安全護欄，使其輸出錯誤的、有害的、甚至是惡意的指令[99,100]。這種攻擊特別危險，因為它可能導致 AI 助手產生有害內容或洩露敏感資訊。

這些層出不窮的對抗性威脅，要求企業在使用 AI 進行安全防護的同時，也必須對其自身的 AI 模型進行持續的、嚴格的壓力測試和安全驗證[81,83]。防禦措施包括對抗性訓練、模型堅固性測試以及多層次的檢測機制[62,94,95,96]。

### ( 三 ) 模型可解釋性的關鍵挑戰

許多先進的、性能卓越的 AI 模型，特別是結構複雜的深度學習模型，在外界看來，其內部決策過程常常如同一個不透明的「黑盒子」（Black Box）[77,79]。我們能夠看到它的輸入是什麼，也能看到它的輸出是什麼，但卻很難清晰地理解它內部的具體決策邏輯[78,79]。

這在網路安全領域，是一個極其嚴重的問題[75,80]。當一個 AI WAF 封鎖了一筆金額巨大的重要商業交易請求時，安全團隊必須能夠清晰地向業務部門解釋「為什麼」這個請求被封鎖，以判斷這是一次正確的防禦，還是一次需要立即修正的誤報[73,74]。缺乏可解釋性，不僅給日常的問題排查和故障修復帶來了巨大困難，在面對如 GDPR（通用數據保護條例）等法規的合規性審計時，也可能造成嚴重的障礙[80]。

因此，整個產業界和學術界目前正在大力推動「可解釋性 AI」（Explainable AI, XAI）的發展[77,78]。其目標就是打開 AI 的「黑盒子」，讓 AI 模型的決策過程變得更加透明、可以被人類理解[77,79]。XAI 技術旨在讓 AI 模型能夠為自己的每一個判斷，提供清晰、可靠的依據和解釋[73,74]，正如前文提到的 Imperva 的 AI Explain 功能，就是這一趨勢在商業產品中的成功嘗試[75]。

主要的可解釋性方法包括 LIME（Local Interpretable Model-agnostic Explanations）和 SHAP（SHapley Additive exPlanations）等技術[74,73]，這些方法能夠為單個預測提供局部解釋，幫助用戶理解模型的決策邏輯。

### ( 四 ) 數據品質與專業人才的雙重瓶頸

AI 模型的性能，從根本上說，高度依賴於其訓練數據的品質和數量[16,17]。一個廣為人知的原則是「垃圾進，垃圾出」（Garbage in, garbage out）。如果用來訓練模型的資料存在偏差、標注錯誤或者無法代表真實世界的威脅，那麼訓練出來的模型不僅無法提供有效的安全防護，甚至可能因為做出錯誤的判斷而帶來新的安全風險[18,19]。然而，在資安領域，要獲取大量、經過準確標注、且能持續更新以反映最新威脅的資料集，本身就是一項巨大且昂貴的挑戰[16,18]。

高品質的網路安全數據集需要滿足多個條件[17,19]：首先是數據的多樣性，必須涵蓋各種

類型的攻擊和正常行為；其次是標注的準確性，需要資深安全專家進行精確的威脅分類；第三是時效性，必須及時更新以反映新興威脅；最後是規模充足性，需要足夠大的數據量來訓練複雜的深度學習模型 [16,26]。

與此同時，能夠真正開發、訓練、部署、維護和解讀這些複雜 AI 模型的跨領域專家——即既精通數據科學和機器學習，又深刻理解網路安全攻防技術的專業人才——在全球範圍內都處於極度短缺的狀態 [10,28]。這種人才瓶頸，也成為了限制 AI 技術在更廣泛的企業和組織中普及應用的一個重要現實因素 [28,69]。

這些專業人才需要具備多重技能：深度的機器學習理論基礎、實際的模型開發經驗、豐富的網路安全知識、以及對業務場景的深刻理解 [68,69]。培養這樣的跨領域專家需要長期的教育投資和實踐積累，這進一步加劇了人才供給的緊張局面 [10,68]。

表 2 系統性地分析了 AI 在網路安全領域應用所面臨的關鍵挑戰與風險因素，包括對抗式機器學習威脅的技術複雜性、模型可解釋性需求的合規壓力、高品質訓練數據獲取的現實困難，以及跨領域專業人才短缺的產業瓶頸。此挑戰分析框架為組織制定 AI 安全技術導入策略時的風險評估與應對規劃提供重要的決策支援，確保技術部署的穩健性與可持續性。

**表 2. AI 在網路安全各領域應用技術架構分析 [57,58,67,81,89]**

| 應用領域 | 主要AI技術 | 核心功能 | 技術優勢 | 面臨挑戰 |
|---|---|---|---|---|
| **威脅情資分析** | 深度學習、圖形演算法 | 海量數據處理、模式挖掘、預測性洞察 | TB 級數據處理、自動關聯分析 | 數據品質依賴 |
| **UEBA行為分析** | 非監督式學習、統計分析 | 正常基線建立、異常行為檢測 | 內部威脅檢測、零時差攻擊識別 | 基線建立時間長 |
| **SOAR自動響應** | 決策樹、規則引擎 | 事件響應自動化、多系統協調 | 毫秒級響應、減少人工干預 | 誤判影響業務 |
| **AI驅動WAF** | 監督式+非監督式學習、RNN | 流量基線、異常檢測、機器人識別 | 適應新威脅、自動策略調整 | 對抗式攻擊威脅 |
| **生成式AI助手** | 大型語言模型 | 自然語言查詢、安全知識增強 | 降低技術門檻、提升效率 | 模型可解釋性 |

## 八、未來發展方向與策略建議

### ( 一 ) 技術演進的重要趨勢

儘管在前行的道路上，仍然面臨著對抗式攻擊、模型可解釋性、數據與人才依賴等種種嚴峻的挑戰，但人工智慧在網路安全領域的應用，其發展步伐依然呈現出不可逆轉的、持續加速的增長趨勢。當我們放眼當今的威脅地平線，看到的是日益自動化、規模化、甚至是由 AI 賦能的智慧化網路攻擊時，我們必須清醒地認識到，僅僅依靠傳統的人力堆砌和靜態規則進行防禦，已無異於在資訊時代的數位戰場上「以卵擊石」。

展望未來，AI 在網路安全領域的應用將會變得更加專門化、更加深入、也更加無處不在。幾個值得期待的發展方向包括：

AI 驅動的自適應行為驗證將結合多因素生物特徵，實現持續性身份認證，不再依賴單次驗證，而是持續監控用戶行為模式。基於多因素生物特徵的持續身份認證將整合指紋、聲

紋、行為模式等多種生物特徵，提供更安全可靠的身份驗證機制。專門用於對抗攻擊者 AI 的「反 AI」策略將開發專門的防禦機制，用於檢測和對抗 AI 驅動的攻擊工具。

### ( 二 ) 產業標準化與生態建設

全球各大網路安全廠商，無論是行業巨頭還是創新型公司，均已投入巨額的研發資源，全力推進 AI 相關安全產品的創新，並積極地與頂尖學術界、研究機構展開深度合作，共同應對 AI 安全本身所面臨的新興威脅，形成一個良性的創新生態。

隨著 AI 安全技術的成熟，建立統一的評估標準、認證體系和最佳實務指引變得日益重要。這包括 AI 模型的安全性評估標準、對抗性攻擊的測試方法、可解釋性的量化指標，以及跨組織的威脅情報共享機制等。聯邦學習技術在威脅情報共享中的應用將使多個組織能夠在不洩露敏感數據的前提下，共同訓練更強大的 AI 安全模型。

### ( 三 ) 決策者的戰略思考與行動建議

對於身處數位轉型浪潮中的現代企業決策者、資訊長（CIO）、資安長（CISO）以及每一位資安專業人士而言，一個清晰的結論已經擺在眼前：將 AI 視為整體安全戰略中不可或缺的核心戰略組件，並持續關注其在安全領域的最新發展，客觀且全面地評估引入 AI 防禦解決方案所能帶來的巨大效益與潛在風險，這已不再是一個可以選擇的選項，而是確保組織的數位資產、商業信譽和持續運營能力，在日益複雜和敵對的全球網路威脅環境中得以生存和發展的必要之舉。

企業在評估和部署 AI 安全解決方案時，應充分考慮技術成熟度、實施成本及長期維護需求。建議採用分階段的部署策略，從風險較低的應用場景開始，逐步擴展到核心業務系統。同時，企業需要建立相應的人才培養和技術更新機制，確保能夠有效管理和維護 AI 安全系統。

## 九、結論

本研究透過系統性的分析與深度的案例研究，全面探討了人工智慧在網路安全領域的應用現況與發展趨勢，特別聚焦於 AI 驅動的網路應用程式防火牆技術。研究結果清晰地表明，AI 技術正在從根本上改變網路安全的防禦哲學與實踐體系，實現了從被動事件響應到主動威脅預測的重要範式轉移。

在威脅檢測與防護效能方面，AI 技術展現了顯著的優勢。透過智慧化威脅情資分析，系統能夠處理 TB 級的安全數據並自動發現隱藏的攻擊模式。UEBA 技術透過建立多維度行為基線，有效識別內部威脅和零時差攻擊的早期階段。AI 驅動的 SOAR 平台將平均應變時間縮短到分鐘級別，真正實現了機器速度的威脅響應。

在 WAF 技術創新方面，AI 系統透過流量行為基線建立實現了異常檢測，不依賴已知攻擊特徵即可識別新型威脅。零時差攻擊預防機制透過學習攻擊序列和意圖，能夠在漏洞公開前就提供保護。智慧化機器人辨識系統準確區分惡意機器人、搜尋引擎爬蟲和真實用戶，並採取差異化應對策略。自適應策略管理功能大幅簡化了規則配置和維護工作。

業界案例驗證了 AI 技術的實際部署價值。Fortinet 的雙層機器學習架構實現了 98% 以上的檢出率和 0.1% 以下的誤報率。AWS 的置信度評分系統提供了直觀的機器人風險評估。Cloudflare 的自然語言操作介面大幅降低了使用門檻。Imperva 的 AI Explain 功能有效橋

接了技術與業務之間的知識鴻溝。Microsoft Security Copilot 展現了生成式 AI 在安全運營中的巨大潛力。

然而，研究也識別出需要持續關注的重要挑戰。對抗式機器學習威脅要求企業對 AI 模型本身進行安全防護。模型可解釋性問題影響了 AI 系統在關鍵決策中的可信度。高品質訓練數據的獲取和跨領域專業人才的短缺成為技術普及的瓶頸。這些挑戰需要學術界、產業界和政府部門的共同努力來解決。

展望未來，AI 在網路安全領域的應用將更加深入和專業化。自適應行為驗證、多因素生物特徵認證、聯邦學習威脅情報共享等新技術將進一步提升防護能力。產業標準化和生態建設將為技術的健康發展提供支撐。

對於企業決策者而言，將 AI 技術納入整體安全戰略已經從可選項轉變為必需品。建議企業採用分階段的部署策略，從低風險應用開始逐步擴展，並建立相應的人才培養和技術更新機制。只有這樣，才能確保組織在日益複雜的網路威脅環境中保持競爭優勢和安全防護能力。

## 參考資料


[1] Akamai Security Research, "Rethinking Defense for Web Apps and APIs: New State of the Internet Report," 2024. [線上]. 可取得 : https://www.akamai.com/blog/security-research/rethinking-defense-web-apps-api-new-soti

[2] M. A. Ferrag, L. Maglaras, S. Moschoyiannis, and H. Janicke, "Deep learning for cyber security intrusion detection: Approaches, datasets, and comparative study," Journal of Information Security and Applications, vol. 50, 102419, 2020.

[3] A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," IEEE Communications Surveys & Tutorials, vol. 18, no. 2, pp. 1153-1176, 2016.

[4] Y. Liu, C. Sarabi, J. Zhang, P. Naghizadeh, M. Karir, and M. Bailey, "Cloudy with a chance of breach: Forecasting cyber security incidents," in Proc. USENIX Security Symposium, 2015, pp. 1009-1024.

[5] D. E. Denning, "An intrusion-detection model," IEEE Trans. Software Engineering, vol. SE-13, no. 2, pp. 222-232, 1987.

[6] S. Axelsson, "Intrusion detection systems: A survey and taxonomy," Technical Report 99-15, Department of Computer Engineering, Chalmers University of Technology, 2000.

[7] P. Garcia-Teodoro, J. Diaz-Verdejo, G. Macia-Fernandez, and E. Vazquez, "Anomaly-based network intrusion detection: Techniques, systems and challenges," Computers & Security, vol. 28, no. 1-2, pp. 18-28, 2009.

[8] C. Kruegel and G. Vigna, "Anomaly detection of web-based attacks," in Proc. ACM Conference on Computer and Communications Security, 2003, pp. 251-261.

[9] W. G. J. Halfond, J. Viegas, and A. Orso, "A classification of SQL-injection attacks and countermeasures," in Proc. IEEE International Symposium on Secure Software Engineering, 2006, pp. 13-15.

[10] Fortinet, "Artificial Intelligence in Cybersecurity," Fortinet Cyber Glossary, 2024. [ 線上 ]. 可取得 : https://www.fortinet.com/resources/cyberglossary/artificial-intelligence-in-cybersecurity

[11] F. Valeur, D. Mutz, and G. Vigna, "A learning-based approach to the detection of SQL attacks," in Proc. Conference on Detection of Intrusions and Malware & Vulnerability Assessment, 2005, pp. 123-140.

[12] M. Roesch, "Snort: Lightweight intrusion detection for networks," in Proc. LISA Conference, 1999, pp. 229-238.

[13] V. Paxson, "Bro: A system for detecting network intruders in real-time," Computer Networks, vol. 31, no. 23-24, pp. 2435-2463, 1999.

[14] K. Scarfone and P. Mell, "Guide to intrusion detection and prevention systems (IDPS)," NIST Special Publication 800-94, 2007.

[15] S. T. Zargar, J. Joshi, and D. Tipper, "A survey of defense mechanisms against distributed denial of service (DDoS) flooding attacks," IEEE Communications Surveys & Tutorials, vol. 15, no. 4, pp. 2046-2069, 2013.

[16] N. Moustafa and J. Slay, "UNSW-NB15: A comprehensive data set for network intrusion detection systems," in Proc. Military Communications and Information Systems Conference, 2015, pp. 1-6.

[17] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set," in Proc. IEEE Symposium on Computational Intelligence for Security and Defense Applications, 2009, pp. 1-6.

[18] L. Dhanabal and S. P. Shantharajah, "A study on NSL-KDD dataset for intrusion detection system based on classification algorithms," International Journal of Advanced Research in Computer and Communication Engineering, vol. 4, no. 6, pp. 446-452, 2015.

[19] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization," in Proc. International Conference on Information Systems Security and Privacy, 2018, pp. 108-116.

[20] G. Creech and J. Hu, "A semantic approach to host-based intrusion detection systems using contiguous and discontiguous system call patterns," IEEE Trans. Computers, vol. 63, no. 4, pp. 807-819, 2014.

[21] H. Debar, M. Dacier, and A. Wespi, "Towards a taxonomy of intrusion-detection systems," Computer Networks, vol. 31, no. 8, pp. 805-822, 1999.

[22] R. Sommer and V. Paxson, "Outside the closed world: On using machine learning for network intrusion detection," in Proc. IEEE Symposium on Security and Privacy, 2010, pp. 305-316.

[23] A. P. Muniyandi, R. Rajeswari, and R. Rajaram, "Network anomaly detection by cascading k-means clustering and C4.5 decision tree algorithm," Procedia Engineering, vol. 30, pp. 174-182, 2012.

[24] M. H. Bhuyan, D. K. Bhattacharyya, and J. K. Kalita, "Network anomaly detection: Methods, systems and tools," IEEE Communications Surveys & Tutorials, vol. 16, no. 1, pp. 303-336, 2014.

[25] C. F. Tsai, Y. F. Hsu, C. Y. Lin, and W. Y. Lin, "Intrusion detection by machine learning: A review," Expert Systems with Applications, vol. 36, no. 10, pp. 11994-12000, 2009.

[26] A. Khraisat, I. Gondal, P. Vamplew, and J. Kamruzzaman, "Survey of intrusion detection systems: Techniques, datasets and challenges," Cybersecurity, vol. 2, no. 1, pp. 1-22, 2019.

[27] Y. Mirsky, T. Doitshman, Y. Elovici, and A. Shabtai, "Kitsune: An ensemble of autoencoders for online network intrusion detection," in Proc. Network and Distributed System Security Symposium, 2018.

[28] Overture Partners, "Emerging AI Trends in Cybersecurity," IT Staffing Resources, 2024. [ 線上 ]. 可取得 : https://overturepartners.com/it-staffing-resources/emerging-ai-trends-in-cybersecurity

[29] S. Potluri and C. Diedrich, "Accelerated deep neural networks for enhanced intrusion detection system," in Proc. IEEE Conference on Emerging Technologies and Factory Automation, 2016, pp. 1-8.

[30] J. Zhang, Y. Chen, and H. Qi, "A survey on deep learning for cybersecurity," Proceedings of the IEEE, vol. 108, no. 9, pp. 1561-1577, 2020.

[31] Gcore, "How AI Enhances WAF/WAAP," Learning Center, 2024. [ 線上 ]. 可取得 : https://gcore.com/learning/how-ai-enhances-waf-waap

[32] K. Leung and C. Leckie, "Unsupervised anomaly detection in network intrusion detection using clusters," in Proc. Australasian Conference on Computer Science, 2005, pp. 333-342.

[33] M. Amer, M. Goldstein, and S. Abdennadher, "Enhancing one-class support vector machines for unsupervised anomaly detection," in Proc. ACM SIGKDD Workshop on Outlier Detection and Description, 2013, pp. 8-15.

[34] L. M. Manevitz and M. Yousef, "One-class SVMs for document classification," Journal of Machine Learning Research, vol. 2, pp. 139-154, 2001.

[35] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," ACM Computing Surveys, vol. 41, no. 3, pp. 1-58, 2009.

[36] M. Goldstein and S. Uchida, "A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data," PloS one, vol. 11, no. 4, e0152173, 2016.

[37] H. Song, Z. Jiang, A. Men, and B. Yang, "A hybrid semi-supervised anomaly detection model for high-dimensional data," Computational Intelligence and Neuroscience, vol. 2017, 8501683, 2017.

[38] L. Ruff, R. Vandermeulen, N. Goernitz, L. Deecke, S. A. Siddiqui, A. Binder, E. Müller, and M. Kloft, "Deep one-class classification," in Proc. International Conference on Machine Learning, 2018, pp. 4393-4402.

[39] A. Graves, A. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, 2013, pp. 6645-6649.

[40] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436-444, 2015.

[41] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. MIT Press, 2016.

[42] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," Neural Computation, vol. 12, no. 10, pp. 2451-2471, 2000.

[43] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," arXiv preprint arXiv:1406.1078, 2014.

[44] A. Vaswani et al., "Attention is all you need," in Proc. Advances in Neural Information Processing Systems, 2017, pp. 5998-6008.

[45] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.

[46] T. Brown et al., "Language models are few-shot learners," in Proc. Advances in Neural Information Processing Systems, 2020, pp. 1877-1901.

[47] Fortinet, "FortiWeb Web Application Firewall," Product Information, 2024. [ 線上 ]. 可取得 : https://www.fortinet.com/products/web-application-firewall/fortiweb

[48] D. Appelt, C. D. Nguyen, L. C. Briand, and N. Alshahwan, "Automated testing for SQL injection vulnerabilities: An input mutation approach," in Proc. ACM SIGSOFT International Symposium on Software Testing and Analysis, 2014, pp. 259-269.

[49] M. Stuttard and M. Pinto, The Web Application Hacker's Handbook: Finding and Exploiting Security Flaws, 2nd ed. Wiley, 2011.

[50] J. Williams and D. Wichers, "OWASP Top Ten 2017: The ten most critical web application security risks," The Open Web Application Security Project, 2017.

[51] B. Sullivan and V. Liu, "Web application security assessment tools," IEEE Security & Privacy, vol. 9, no. 4, pp. 32-41, 2011.

[52] G. Wassermann and Z. Su, "Static detection of cross-site scripting vulnerabilities," in Proc. International Conference on Software Engineering, 2008, pp. 171-180.

[53] Y. Huang, F. Yu, C. Hang, C. Tsai, D. Lee, and S. Kuo, "Securing web application code by static analysis and runtime protection," in Proc. International Conference on World Wide Web, 2004, pp. 40-52.

[54] L. Vieira, "Web application firewall: Your first line of defense," Information Security Journal: A Global Perspective, vol. 16, no. 5, pp. 233-241, 2007.

[55] AWS, "AWS WAF Bot Control Protects Against Distributed Proxy-Based Attacks," AWS News, 2023. [ 線上 ]. 可取得 : https://aws.amazon.com/about-aws/whats-new/2023/09/aws-waf-bot-control-protects-against-distributed-proxy-based-attacks/

[56] P. E. Ayres, H. Sun, S. Chao, and L. Ristenpart, "SoK: Understanding BEC scams," in Proc. IEEE Symposium on Security and Privacy, 2021, pp. 1481-1495.

[57] B. Biggio and F. Roli, "Wild patterns: Ten years after the rise of adversarial machine learning," Pattern Recognition, vol. 84, pp. 317-331, 2018.

[58] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in Proc. IEEE European Symposium on Security and Privacy, 2016, pp. 372-387.

[59] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," arXiv preprint arXiv:1412.6572, 2014.

[60] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," arXiv preprint arXiv:1607.02533, 2016.

[61] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, "Ensemble adversarial training: Attacks and defenses," arXiv preprint arXiv:1705.07204, 2017.

[62] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," arXiv preprint arXiv:1706.06083, 2017.

[63] Cloudflare, "Making WAF ML Models Go Brrr," Cloudflare Blog, 2024. [ 線上 ]. 可取得 : https://blog.cloudflare.com/making-waf-ai-models-go-brr/

[64] Cloudflare, "AI Everywhere with the WAF Rule Builder Assistant," Cloudflare Blog, 2024. [ 線上 ]. 可取得 : https://blog.cloudflare.com/bringing-ai-to-cloudflare/

[65] B. Nelson, M. Barreno, F. J. Chi, A. D. Joseph, B. I. P. Rubinstein, U. Saini, C. Sutton, J. D. Tygar, and K. Xia, "Exploiting machine learning to subvert your spam filter," in Proc. USENIX Workshop on Large-Scale Exploits and Emergent Threats, 2008, pp. 1-9.

[66] M. Barreno, B. Nelson, A. D. Joseph, and J. D. Tygar, "The security of machine learning," Machine Learning, vol. 81, no. 2, pp. 121-148, 2010.

[67] L. Huang, A. D. Joseph, B. Nelson, B. I. Rubinstein, and J. D. Tygar, "Adversarial machine learning," in Proc. ACM Workshop on Security and Artificial Intelligence, 2011, pp. 43-58.

[68] Darktrace, "How AI is Transforming Cybersecurity Practices," Security Blog, 2024. [ 線上 ]. 可取得 : https://www.darktrace.com/blog/how-ai-is-transforming-cybersecurity-practices

[69] CrowdStrike, "CrowdStrike Named 2025 Gartner Magic Quadrant Leader," Press Release, 2024. [ 線上 ]. 可取得 : https://www.crowdstrike.com/en-us/press-releases/crowdstrike-named-2025-gartner-magic-quadrant-leader-endpoint-security/

[70] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785-794.

[71] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.

[72] C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, no. 3, pp. 273-297, 1995.

[73] S. M. Lundberg and S. Lee, "A unified approach to interpreting model predictions," in Proc. Advances in Neural Information Processing Systems, 2017, pp. 4765-4774.

[74] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?: Explaining the predictions of any classifier," in Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135-1144.

[75] Imperva, "Bridging the Security Knowledge Gap: Introducing AI Explain," Imperva Blog, 2024. [ 線上 ]. 可取得 : https://www.imperva.com/blog/bridging-the-security-knowledge-gap-introducing-ai-explain-for-imperva-cloud-waf/

[76] Microsoft, "Microsoft Security Copilot," Microsoft Security, 2024. [ 線上 ]. 可取得 : https://www.microsoft.com/en-us/security/business/ai-machine-learning/microsoft-security-copilot

[77] D. Gunning, "Explainable artificial intelligence (XAI)," Defense Advanced Research Projects Agency (DARPA), 2017.

[78] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," IEEE Access, vol. 6, pp. 52138-52160, 2018.

[79] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," Nature Machine Intelligence, vol. 1, no. 5, pp. 206-215, 2019.

[80] B. Goodman and S. Flaxman, "European Union regulations on algorithmic decision-making and a 'right to explanation'," AI Magazine, vol. 38, no. 3, pp. 50-57, 2017.

[81] ISACA, "Combating the Threat of Adversarial Machine Learning to AI-Driven Cybersecurity," Industry News, 2025. [ 線上 ]. 可取得 : https://www.isaca.org/resources/news-and-trends/industry-news/2025/combating-the-threat-of-adversarial-machine-learning-to-ai-driven-cybersecurity

[82] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in Proc. IEEE Symposium on Security and Privacy, 2017, pp. 39-57.

[83] A. Athalye, N. Carlini, and D. Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," in Proc. International Conference on Machine Learning, 2018, pp. 274-283.

[84] F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart, "Stealing machine learning models via prediction APIs," in Proc. USENIX Security Symposium, 2016, pp. 601-618.

[85] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in Proc. ACM SIGSAC Conference on Computer and Communications Security, 2015, pp. 1322-1333.

[86] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in Proc. IEEE Symposium on Security and Privacy, 2017, pp. 3-18.

[87] AWS, "How to Improve Visibility into AWS WAF with Anomaly Detection," AWS Security Blog, 2024. [ 線上 ]. 可取得 : https://aws.amazon.com/blogs/security/how-to-improve-visibility-into-aws-waf-with-anomaly-detection/

[88] AWS, "How to Manage AI Bots with AWS WAF and Enhance Security," AWS Blog, 2024. [ 線上 ]. 可取得 : https://aws.amazon.com/blogs/networking-and-content-delivery/how-to-manage-ai-bots-with-aws-waf-and-enhance-security/

[89] T. Gu, B. Dolan-Gavitt, and S. Garg, "BadNets: Identifying vulnerabilities in the machine learning model supply chain," arXiv preprint arXiv:1708.06733, 2017.

[90] Y. Liu et al., "Trojaning attack on neural networks," in Proc. Network and Distributed System Security Symposium, 2018.

[91] X. Chen, C. Liu, B. Li, K. Lu, and D. Song, "Targeted backdoor attacks on deep learning systems using data poisoning," arXiv preprint arXiv:1712.05526, 2017.

[92] A. Turner, D. Tsipras, and A. Madry, "Label-consistent backdoor attacks," arXiv preprint arXiv:1912.02771, 2019.

[93] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, "How to backdoor federated learning," in Proc. International Conference on Artificial Intelligence and Statistics, 2020, pp. 2938-2948.

[94] B. Wang et al., "Neural cleanse: Identifying and mitigating backdoor attacks in neural networks," in Proc. IEEE Symposium on Security and Privacy, 2019, pp. 707-723.

[95] K. Liu, B. Dolan-Gavitt, and S. Garg, "Fine-pruning: Defending against backdooring attacks on deep neural networks," in Proc. International Symposium on Research in Attacks, Intrusions, and Defenses, 2018, pp. 273-294.

[96] Y. Gao, C. Xu, D. Wang, S. Chen, D. C. Ranasinghe, and S. Nepal, "STRIP: A defence against trojan attacks on deep neural networks," in Proc. Annual Computer Security Applications Conference, 2019, pp. 113-125.

[97] F. Wallace, R. S. S. Kumar, and E. Hemberg, "Red teaming language model detectors with language models," arXiv preprint arXiv:2305.19713, 2023.

[98] N. Perez, J. Tack, S. Choi, P. Ribeiro, A. F. T. Martins, and J. Shin, "Ignore previous prompt: Attack techniques for language models," arXiv preprint arXiv:2211.09527, 2022.

[99] Y. Liu, G. Deng, Z. Xu, Y. Li, Y. Zheng, Y. Zhang, L. Zhao, T. Zhang, and Y. Liu, "Jailbreaking ChatGPT via prompt engineering: An empirical study," arXiv preprint arXiv:2305.13860, 2023.

[100] A. Wei, N. Haghtalab, and J. Steinhardt, "Jailbroken: How does LLM safety training fail?," in Proc. Conference on Neural Information Processing Systems, 2023.