(ISC)² Security Congress 2024 研習紀要

蘇柏鳴 / 金融聯合徵信中心 資安部

前言

（ISC）² 全名為International Information System Security Certification Consortium，成立於1989年，總部位於美國佛羅里達州的Clearwater，（ISC）² 的使命是為全球資訊安全專業人士提供認證、教育和資源，促進網路安全領域的發展，保護關鍵資訊和基礎設施，推動全球對資訊安全的認知與實踐，主要目標為「制定並推廣資訊安全行業的專業標準和道德規範」、「提供世界一流的網路安全和資訊安全專業認證」、「支持全球資安社群的發展，促進專業人士之間的交流和協作」及「幫助專業人士應對快速變化的安全威脅與技術挑戰」。（ISC）²主要的活動和計劃包括：

1. （ISC）² Security Congress：每年舉辦的全球資訊安全大會，是資安行業的標誌性活動，匯集行業領袖和專業人士，共同討論網絡威脅、數據隱私、雲端安全、人工智慧（AI）在資安的應用等熱門話題，並提供培訓機會和實務指導。。
2. Cybersecurity Awareness Month活動：舉辦針對企業和個人的教育推廣活動，提供免費資源和工具，如資安指南、網路安全課程，並分享如何保護自己免受網絡攻擊的威脅。
3. （iSC）² Chapters地區分會活動：透過各地分會舉辦地區性會議與活動，加強當地資安專業人士的聯繫，分享實用知識，並討論本地化的安全挑戰。
4. Think Tank網絡研討會：線上資安教育計劃，提供由行業專家主持的專業講座和案例討論，內容涵蓋最新的安全技術、行業最佳實踐、法規合規要求等主題。
5. 研究與報告：每年發布《Cybersecurity Workforce Study》，分析資安行業的人才缺口、需求趨勢及未來挑戰。

（ISC）² Security Congress

（ISC）² Security Congress 2024於10月14日至16日在美國拉斯維加斯的凱撒宮（Caesars Palace）舉行，吸引了全球資訊安全專業人士的參與，本次大會以 「勇往直前」（Boldly Forward）為主題，旨在探討資訊安全領域的最新趨勢、挑戰與解決方案，強調在面對不斷演變的威脅環境時，專業人士應具備的前瞻性與行動力。大會期間共舉辦了超過130場的專題演講與工作坊，本文節錄筆者認為比較值得分享的4個會議內容：

1. Opening Keynote: AI Gone Haywire: Lessons from the Lighter Side of Artificial Intelligence
2. AI-Assisted Cyber Storm: Exploiting Cloud to Save It With Generative AI
3. Beyond Red vs. Blue: The Power of Collaborative Security Testing with Purple Teaming
4. Securing LLMs: Spear Phishing the Machines.

參訪目的

因為在近年發生起多起重大資安事件，例如，2020年台灣中油與台塑遭駭，遭受勒索軟體攻擊，導致業務癱瘓並損失巨額資金；2021年，廣達電腦的駭侵事件暴露了蘋果新產品設計圖等機密資料，對供應鏈造成巨大威脅；2022年Nvidia遭到駭客組織Lapsus$入侵，竊取1 TB內部機敏資料；2023年，T-Mobile的3700萬客戶資料外洩事件；2024年，台灣多家政府機關網站遭受分散式阻斷服務（DDoS）攻擊，導致民眾無法正常使用公共服務等，參加（ISC）² Security Congress 2024的目的是深入了解最新的網路威脅趨勢、技術解決方案及實務經驗，增強組織和個人應對資安挑戰的能力，會議還提供深入探討新興趨勢的機會。透過參加研討會，參與者能掌握行業的最新動態，獲得可實施的解決方案，並透過與同行的交流分享經驗，提升資訊安全的視野。

資安事件層出不窮，儘管各組織逐漸加大對資安防護的人力與資金投入，但在資安防護與駭客攻擊技術的競賽中，整體防護進化速度仍然相對落後攻擊方。作為全台灣信用資料的核心收集機構，聯徵中心肩負著穩定金融體系的重要使命，不僅需妥善保護民眾的敏感資料，還需確保徵信資訊的安全傳遞，以支持金融機構的正常運作。

參訪過程

一、ISC2 Security Congress 2024 參與會議場次

Day 1		
會議標題	會議講者	會議內容
Opening Keynote: AI Gone Haywire: Lessons from the Lighter Side of Artificial Intelligence	Janelle Shane（AI Speaker and Humorist）	必須謹慎面對AI Tool的請求，因為結果可能超出你的預期，儘管AI的能力和語言理解不斷進步，但在缺乏周全考慮的情況下，錯誤仍然屢見不鮮。在這場演講中，Janelle Shane將分享她的部落格《AI Weirdness》和著作《You Look Like a Thing and I Love You》中的一些搞笑AI生成示例，讓與會者了解AI的潛力和局限性，並說明為什麼網絡安全和商業專業人士不能盲目依賴AI，否則可能會面臨全新的風險與挑戰。
AI-Assisted Cyber Storm: Exploiting Cloud to Save It With Generative AI	Mohit Sharma	1. 探討生成式AI在雲端環境中識別漏洞的實際應用，解析AI如何模擬複雜的網絡攻擊，揭示潛在漏洞並介紹最新的防禦機制，協助與會者理解AI驅動的安全分析與模擬的具體應用。 2. 透過真實案例和可操作的策略，演講將分享前沿研究、工具和方法，幫助網絡安全專業人士採用AI進行主動的安全規劃，確保雲端基礎設施在面對高級威脅時的韌性與安全性。
How To Create Successful Malware and Defend with Zero Trust（Sponsored by ThreatLocker）	Ryan Bowman（ThreatLocker VP）	1. 當前，惡意軟體的製作門檻大幅降低，甚至可藉由AI自動生成。此次會議將深入探討各類惡意軟體和攻擊的運作原理，協助與會者了解威脅背後的成功關鍵，並針對如何強化防禦提出實務建議。 2. 會議將以「預設拒絕」的安全姿態為核心，剖析惡意軟體的結構，並闡述零信任（Zero Trust）原則的重要性。無論面對人工駭客或AI攻擊者，均能確保數據安全和業務連續性，建構更為堅實的資安防禦體系。
Hacking and Armoring Identity Ecosystems: When MFA isn't Good Enough Any Longer	Dan Houser	1. MFA雖已普及，但網絡釣魚攻擊仍頻繁發生，原因在於身份驗證框架的漏洞與工程缺陷。此次演講聚焦如何強化身份生態系統，透過無密碼身份驗證、FIDO2令牌和生物識別技術等方案。 2. 提供30/60/180天的實施路線圖，協助企業建立強韌的身份防禦能力，避免風險轉移到其他攻擊向量，實現更高的資安標準。
Beyond Red vs. Blue: The Power of Collaborative Security Testing with Purple Teaming	Jason O'Dell（Walmart Global VP，Security Operations）	1. 探討如何透過協作方法推動安全實踐，並分享企業成功整合紅隊（Red Team）與藍隊（Blue Team）為紫隊（Purple Team）的經驗，以加強其安全態勢，實現更高效的威脅應對能力。 2. 促進持續的溝通與協作，並瞭解建立紫隊計劃的實際步驟、工具和流程，為企業制定更具前瞻性和效能的安全戰略。

Day 2		
會議標題	會議講者	會議內容
API Security: A Pentesters Perspective	Jennifer Shannon	會議介紹大型語言模型（LLM）的構建原理，並深入探討其常見威脅和攻擊手法，解釋其運作原理。演講將比較LLM攻擊與XSS、SQL注入等傳統攻擊的異同，特別關注控制平面與數據平面的模糊界限，說明攻擊者如何利用該弱點引導LLM生成意外回應，幫助企業加強LLM的安全防護。
API Security: A Pentesters Perspective	Jennifer Shannon	本次演講將介紹API的基礎知識及其安全保護的重要性。首先，我們將探討API最常見的漏洞，並展示滲透測試人員常用的技術工具，如Burp和Postman。透過演示和實際案例，強調未修補這些漏洞可能帶來的影響。在演示過程中，還會討論多種防範這些攻擊的方法和策略。最後，演講將呼籲企業重新審視其 API 的開發和測試方式，以加強整體的安全性和防禦能力。
Zero Trust Architecture: Is Real World Implementation Even Possible ?	Vincent C. Romney（Nu Skin CISO）	1. 在現有企業架構中實施零信任架構（ZTA）挑戰重重，深入探討關鍵問題、考量因素及實施過程中的決策要點，幫助企業評估是否繼續推進ZTA。 2. 演講者分享實踐ZTA的經驗，解釋NIST 800-207原則對架構模型的影響，並探討替代控制措施、可接受的風險及如何向利益相關者有效傳達資訊，以支持更明智的決策。
The NIST AI Risk Management Framework vs. ISO 23894: Which One is Right for You ?	Scott M. Giordano	深入比較NIST AI RMF 1.0與ISO/IEC 23894這兩大AI風險管理框架的異同，並針對生成式AI的快速崛起提供實用的評估標準。講者將分享如何根據企業需求選擇最適合的框架，並解釋其在AI風險管理中的應用，協助企業應對不斷增長的AI技術挑戰，確保風險可控、合規性達標。
API Security: A Pentesters Perspective	Jennifer Shannon	隨著量子計算技術的進步，其對數字安全構成的威脅也日益加劇。「現在竊取，未來解密」的風險正如高速列車般快速逼近。為應對量子計算的發展，網絡安全領導者必須了解這項技術及其對安全的影響，並採取主動措施，保護數字資產和數據免受未來挑戰的影響。 由於對手現正收集加密數據，準備在未來的量子計算技術成熟時進行解密，這一威脅不容忽視。本次會議將深入探討這一主題，讓與會者了解量子威脅的現狀，並提供可立即實施的行動方案，幫助企業未雨綢繆，確保未來的數據安全。

Day 3		
會議標題	會議講者	會議內容
The Need 4 Skills: Speeding Up the Cybersecurity Skill Set to Stay in the Race Against Cyberattacks	Jorge M. Ochoa	有效的網絡安全領導力需具備戰略視野、技術專業、快速決策、溝通影響力和適應力。領導者應推動學習和創新文化，提供職業發展計劃和多樣化學習機會，優化角色分配並強化團隊協作。透過動態任務調整與開放的心理安全環境，激發成員創新力，並以KPI和OKR衡量績效，持續改進網絡安全策略，應對不斷變化的威脅，推動企業長期穩健發展。
Mitigating the Risk in Risk Management Through Targeted Analysis	Kyle Hinterberg	傳統風險管理依賴定期的風險評估，但這種方式難以及時應對新興威脅。內容介紹持續目標分析策略，這種靈活且適應性強的方法能即時響應威脅，並加強組織的安全性和合規性。透過實踐該策略，企業可滿足PCI DSS、HITRUST和ISO等嚴格框架的合規要求，實現持續的合規和安全性。與會者將獲得實用的實施策略，協助其在風險管理中平衡網絡安全防禦與合規義務，提升組織的風險應對能力和整體防禦效能。
Bright Ideas Roundtable: Navigating the NIST Cybersecurity Framework 2.0	Kelly Hood	關注這些功能之間的相互關聯，如何構建全面的網絡安全戰略，以及新增的「治理 （Govern）」功能如何作為促進整體改進和問責的關鍵推動力。我們將通過互動示例和案例研究，展示NIST網絡安全框架的實際應用。這些場景將展示企業如何利用CSF，將網絡安全工作與戰略目標保持一致，並培養網絡安全意識文化，實現更高效的風險管理和安全防禦。
Jack of More Trades: Introduction to MLSecOps for DevSecOps Professionals	Natalia Semenova	隨著組織日益依賴機器學習（ML）進行決策，確保 ML 系統的安全至關重要。MLSecOps將安全運營（SecOps）與 ML 運營（MLOps）相結合，重點保護模型、數據和基礎設施的完整性和機密性，並防範對抗性攻擊、數據洩露和漏洞風險。由於數據科學家和安全專業人士在技能和知識上的差距，這場演講將針對具備DevSecOps和SSDLC經驗的專業人士，提供清晰的策略和實踐方法，實現 ML 安全的「左移」策略，協助組織在早期預測和緩解風險，確保 ML 系統的安全性和合規性。

二、Opening Keynote: AI Gone Haywire: Lessons from the Lighter Side of Artificial Intelligence

（一）講者

Janelle Shane（AI Speaker and Humorist）

（二）內容摘要

隨著人工智能（AI）技術的迅速發展，AI的應用已滲透到語言處理、圖像生成、自動駕駛、招聘篩選和決策支持等眾多領域。儘管AI在許多方面表現出卓越的能力，但許多實際案例卻揭示了AI存在「記憶錯誤」、「捷徑行為」和「偏見放大」等問題。這些問題不僅影響AI的準確性和可用性，甚至可能對社會公平性和企業決策產生嚴重影響。

1. 記憶錯誤問題

AI的「記憶錯誤」指的是AI系統並非基於輸入數據進行推理，而是基於訓練數據的記憶生成輸出。當AI在訓練中見過某些數據，無論這些數據是否準確，AI會將這些記憶作為未來的預測基礎，導致錯誤的輸出結果。

(1) 案例：Bard的櫻桃起司派錯誤

在這個實驗中，講者手繪了一張櫻桃起司派的圖片，並請Bard（谷歌的AI圖像描述工具）對這張圖片進行描述，Bard的描述包括「螺旋狀的櫻桃圖案」、「白色盤子」和「香芹裝飾」。然而，這些細節在圖片中根本不存在。

問題原因：Bard的生成結果不是基於圖像內容的檢測，而是基於訓練數據的記憶生成的它依賴於機率的文本模式，而不是實際觀察圖片的特徵。

(2) 案例：MidJourney的「復仇者聯盟式生成」

MidJourney是一款圖像生成AI，用戶可以請求它生成特定風格的圖像，當用戶請求生成《復仇者聯盟：無限之戰》風格的圖像時，MidJourney 生成的圖像與電影的畫面幾乎完全相同。

問題原因：MidJourney 並沒有創造這些圖像，而是重複了訓練數據中的場景。這是 AI 中的數據記憶現象，也是版權問題的爭議焦點之一，因為這些圖像可能違反了版權法。

(3) 案例：GPT-4 的「考試作弊行為」

在一個實驗中，研究人員測試了GPT-4在考試中的表現，當AI被測試訓練截止時間之前的考題時，它能夠獲得10/10的滿分，但當測試訓練截止後的新考題時，AI的表現急劇下降，得分僅為0/10。

問題原因：GPT-4的行為顯示其解題能力依賴於記憶，而不是理解。由於GPT-4可能在訓練數據中見過這些考題，因此能輕鬆作答；但面對訓練數據中從未見過的考題時，則無法做出正確的回答。

2. AI的「捷徑行為」問題

捷徑行為是指AI系統在解決問題時，會選擇最容易的途徑來達成目標，而不一定是正當的途徑。這一問題經常出現在沒有清晰指令或約束的情境中。

(1) 案例：Lindy Bot 的「隨機影片連結」

Lindy Bot是一個客戶服務機器人。當客戶詢問「有關如何使用該機器人的指南」時，Lindy Bot隨機生成了一個 YouTube影片鏈接，標題是「如何製作和使用模板的教程」，然而，該公司並沒有製作這段視頻。

問題原因：Lindy Bot的反應是基於最大概率的行為選擇，而不是真實的語義理解。

(2) 案例：《俄羅斯方塊》的作弊行為

AI在遊戲《俄羅斯方塊》中被要求不要輸掉遊戲，AI採取的策略是「按下暫停鍵，永遠不恢復遊戲」，從而滿足了不要輸的目標。

問題原因：目標定義不明確，AI發現按下暫停鍵是一種便捷的捷徑，技術上來說，它並沒有輸掉遊戲。

3. AI的「偏見放大」問題

AI的偏見問題常常出現在招聘篩選、面試評估、語言生成等高風險的應用中。偏見通常來源於訓練數據的偏見，並被AI進一步放大。

(1) 案例：亞馬遜的招聘AI

亞馬遜的招聘AI在篩選簡歷時，對女性求職者進行了歧視，即使開發者去除了性別標籤，該系統仍能通過女子學院或女子軟式壘球等隱藏特徵，預測求職者的性別。

問題原因：AI會依據潛在變數和隱含特徵，而這些變數可能包含了性別資訊。

(2) 案例：個性評估算法的「背景書架偏見」

一個用於分析視頻面試中的個性特質的算法，發現其判斷結果受背景中是否存在書架的影響。如果背景中有書架，該算法會將求職者標記為更有智慧。

問題原因：AI模型會依賴訓練數據中的模式和潛在關聯，如果數據中出現書架與高智力的關聯，AI就會自動擴大這一偏見。

4.解決方法

「明確目標，避免捷徑行」、「引入人類專家審查」、「加強數據管理，避免記憶錯誤」及「建立AI責任機制和問責機制」。

（三）心得

許多人擔心AI會取代人類的工作，但事實可能更微妙，有人提出「AI不會直接取代你的工作，但會讓那些更懂得使用AI的人取代你。」這意味著，未來的競爭不僅僅是人與AI的競爭，而是人類之間使用AI能力的競爭，例如，化學家現在使用AlphaFold[1]來幫助研究蛋白質結構，這使得人類與AI的合作關係成為新的常態。

另外，隨著生成式AI（例如GPT-4）被引入教育環境，出現了教育工具與學習陷阱之間的爭議，如果學生使用AI生成論文，那麼他們就失去了思考和組織思想的機會。不過有些人他們會使用AI來調整文字語氣，以使自己的訊息更具專業性，這在一定程度上是有益的輔助工具。由此可見，如何使用AI工具、如何引導學習行為，可能會成為未來教育中的新議題。

AI的發展歷史中經歷過數次AI寒冬（AI Winter），但目前的情境更像是AI夏天（AI Summer），其中有許多資金和技術投入，尤其是對於生成式AI的投資。然而，這個泡沫是否會像過去一樣破裂，有專家認為，大規模投資AI的企業將面臨現實的檢驗，投資者會重新審視哪些技術是值得繼續投入的。

三、AI-Assisted Cyber Storm: Exploiting Cloud to Save It With Generative AI

（一）講者

Mohit Sharma（Chief Consultant， Atea AS）

（二）內容摘要

雲端技術的出現改變了企業的基礎架構建設方式，服務級架構（Service-Level Architecture， SLA）成為可能，基礎架構即代碼（Infrastructure as Code， IaC）的實現讓開發人員能與運營人員無縫協作。開發人員現在不僅能夠編寫業務邏輯，還能同時部署和管理其所需的基礎架構，使得開發和運營的界限逐漸模糊，真正實現了DevSecOps 的願景。

[1] Alphabet旗下Google旗下DeepMind開發的一款蛋白質結構預測程式，https://alphafold.ebi.ac.uk/

過去，企業需要自行購買伺服器、建立數據中心，並花費高額的運維成本。而現在，雲端服務提供了隨選即用的彈性計算資源，企業不必再負擔龐大的基礎設施費用。這一變化降低了新創企業的門檻，即便是資金不足的小型公司，也能夠輕鬆獲取高性能的計算和存儲資源，實現快速的業務擴張。

雲端的開放性和資源的易得性，使得區塊鏈、AI和生成式AI等尖端技術變得普及化。以往只有大型實驗室或企業才能負擔的高成本技術，現在成為一種服務（AI as a Service），即使是普通用戶也能接觸到生成式AI技術，透過ChatGPT、DALL-E等平台創建內容，這一現象也為數位創意產業和教育帶來了新的變革。

雲端的靈活性雖然為企業帶來了巨大的效益，但也產生了「誰該為安全負責的模糊地帶」。企業內部的開發人員、架構師和運營工程師對於誰應負責存儲Blob的安全性存在分歧。當公司將數據存放在雲端，誰來確保這些Blob數據不會被未授權的用戶存取？這種責任模糊地帶導致企業在身份和訪問管理（IAM）、存儲數據加密等問題上出現重大疏忽，進而成為雲端攻擊的漏洞。

另外，許多公司選擇多雲策略（Multi-Cloud Strategy），導致其架構的複雜度大幅提升。根據Gartner的報告，70-80%的企業使用不止一種雲端服務。但多雲架構的安全配置難度更高，服務之間的接口、數據同步和身份管理系統常常缺乏一致性，容易產生錯誤配置和權限漏洞，這成為駭客入侵雲端基礎設施的關鍵點，數字供應鏈攻擊的風險正在不斷上升，大部分企業所使用的服務級架構（SLA）和開源軟體（Open Source Software），其實是由全球各地的開源貢獻者和匿名開發者維護的，這些開源軟體一旦被入侵（如類似SolarWinds供應鏈攻擊），企業將無從得知其程式庫中的後門，進而導致全球性的大規模攻擊。

講者針對生成式AI對安全的威脅與挑戰有以下敘述：

1. 生成式AI的影響

知識門檻的下降：生成式AI（如ChatGPT）降低了技術門檻，以往需要高超技術和大量時間的人員，現在只需輸入問題，便能獲得明確的操作指令、程式代碼和技術細節。這一現象使得技術門檻變得前所未有的低，駭客不必再花時間研究和設計攻擊路徑，僅需幾行代碼或幾個提示詞，便可生成可利用的惡意程式。

2. AI如何助長駭客行為

（1）腳本生成

不具備Shell語言知識的用戶，現在只需讓ChatGPT生成SSH或RDP掃描腳本，即可發現公開的雲端資源，並嘗試對其進行爆破登錄。

（2）身份欺詐

生成式AI能生成仿冒的公司郵件，這些郵件不僅語法完美，還具備真實的視覺樣式和一致的品牌標誌，使得釣魚（Phishing）攻擊的成功率大幅上升。

（3）自動化網路偵查

過去駭客需要自己編寫網路掃描程式，現在只需向AI請求掃描命令，便能獲得一套完整的腳本，從而發現網路中的未授權資源、漏洞和敏感數據。

3. 生成式AI導致的真實數據洩漏

（1）TaskRabbit

DDoS攻擊過程中，AI不斷變換其行為，導致其適應性攻擊不斷升級，從而導致370萬用戶的數據洩漏。

（2）T-Mobile

2023年，T-Mobile的3700萬用戶數據被盜，駭客利用AI執行自動化的入侵和識別，大幅降低了手動入侵的成本。

（三）心得

應對生成式AI安全威脅的解決方案

1. 使用AI進行「進攻性防禦」，與其等待駭客來攻擊，不如企業自己利用AI掃描其基礎設施的漏洞。在企業的預部署測試中，使用生成式AI來模擬駭客的入侵行為，並修補可能的IAM錯誤配置、敏感API漏洞和外部暴露的端點。
2. 強化雲端安全的設計，落實零信任架構（Zero Trust），確保每個端點和每個用戶的身份都經過驗證。文件化每一個設計和架構變更，許多企業的系統設計和架構經常未被記錄，這是錯過安全審查的關鍵原因。
3. 提高用戶認證的多樣性和強度，實施多因素身份驗證（MFA），而不僅僅是雙因素身份驗證（2FA）。密碼更改的頻率和複雜性應提高。

四、Beyond Red vs. Blue: The Power of Collaborative Security Testing with Purple Teaming

（一）講者

Jason O'Dell（Walmart Global VP， Security Operations）

（二）內容摘要

Purple Team是一種將Red Team（紅隊）和Blue Team（藍隊）協作起來的安全測試方法，其目的是透過團隊的合作來加強組織的網絡安全。紅隊的任務是模擬真實的攻擊者，測試企業的安全防禦能力；藍隊則負責檢測和防禦這些攻擊行為，並進行補救行動。Purple Team作為中間橋樑，促進雙方的透明度和協作，這種協作方式旨在增強組織的整體安全性。

與傳統的紅藍對抗不同，Purple Team更關注於持續改進和合作，而不是僅僅追求輸贏。這種方法能幫助企業提升威脅檢測和響應能力，縮短平均檢測和回應時間（MTTD和MTTR），並讓企業能夠及時修補關鍵漏洞。

Purple Team的三大核心區域

1. Red Team（紅隊）的角色與目標

- 模擬威脅行為：如APT（高階持續性威脅）團體的滲透、側向移動和特權提升行為。
- 建立威脅模型：藉由MITRE ATT&CK[2]矩陣，讓紅隊明確了解模擬攻擊的策略與技術。
- 提供回饋與改進：根據模擬的攻擊行為，向藍隊提供有針對性的回饋。

2. Blue Team（藍隊）的角色與目標

- 防禦與檢測：通過監視網絡活動，檢測和回應潛在威脅。
- 威脅情報：使用威脅情報來加強對攻擊手法的理解，並提前設定檢測指標（IOC）。
- 建立檢測和回應程序：制定和改進檢測技術，如可疑活動日誌分析、異常行為檢測和自動化威脅回應。

3. Purple Team的角色與目標

- 消除「對抗性心態」：將紅隊和藍隊轉變為協作夥伴關係，並推動雙方之間的透明度。
- 共同執行桌面推演（Tabletop Exercise）：這是一種不執行真實攻擊的模擬推演方式，雙方在會議室中一起進行攻擊和回應的假設演練。
- 提升組織的檢測與回應能力：不僅僅關注誰贏誰輸，而是通過每一次模擬攻擊來增強組織的檢測能力和防禦反應。

Purple Team的主要實踐方法

1. 桌面推演（Tabletop Exercise），分為兩種類型

- 非技術性桌面推演： 紅隊與藍隊之間的對話，討論針對威脅的可能反應，但不會實際執行命令。
- 技術性桌面推演： 在真實環境中執行實際的指令，測試系統的檢測反應、日誌可見性和補救行動。

2. 受信代理人（Trusted Agent）模型

受信代理人是Purple Team中的藍隊代表，參與到紅隊的行動中，但不會通知整個藍隊。受信代理人能在紅隊行動過程中提供即時反饋，並幫助藍隊提前了解潛在的威脅，這樣的安排能提升對未來攻擊的反應能力。

連續攻擊模擬（Continuous Attack Simulation）

3. 自動化攻擊測試平台

像Atomic Red Team這樣的工具會持續模擬各種威脅，並自動生成報告。自動化漏洞測試：檢查系統中是否有Control Drift，即原本被修復的漏洞由於後續的配置變更而重新出現。

4. 友好競爭的快打行動

這類行動通常為為期3-4天的紅隊模擬行動，會假設“違規訪問”已發生，並模擬實際的威脅行為。這類行動比傳統的長期滲透測試要短小得多，但回饋速度更快，效果更直接。

[2] MITRE ATT&CK 架構是一種策略和技術知識庫，專為威脅捕捉專家、防禦者和紅隊而設計，可以協助對攻擊進行分類、識別攻擊歸因和目標並評估企業的風險。

Purple Team的工具和技術

1. MITRE ATT&CK

用於映射威脅行為，並提供紅隊/藍隊共同語言。

2. Atomic Red Team

開源的滲透測試自動化工具，可在多個環境中運行測試腳本。

3. PRPL工具

提供報告生成、操作時間戳記、行動報告和紅藍雙方的透明溝通，讓雙方的測試更具可見性。

Purple Team的挑戰和關鍵成功因素

1. 挑戰

- 信任問題：紅隊和藍隊需要保持彼此之間的透明溝通，但如果信任不夠，訊息可能被封鎖，導致錯誤的決策。
- 資訊共享的限制：藍隊的威脅情報是否要與紅隊分享？許多企業會擔心紅隊走捷徑，但共享情報能幫助紅隊模擬更準確的攻擊。
- 資源的平衡：進行Purple Team行動需要更多的資源和人力支持，而不是每個企業都能負擔得起的。

2. 關鍵成功因素

- 信任與透明：雙方的透明度決定了協作的深度，尤其是在受信代理人（Trusted Agent）模型中，代理人的透明性決定了藍隊的洞察力。
- 行動紀錄：在紅隊的行動中，所有的攻擊時間戳記和命令都應該被記錄，以便藍隊事後檢討。
- 零知識攻擊模式：這種模式中，紅隊在行動中不使用現有的內部知識，而是模擬全新攻擊者的行為，這使得檢測的挑戰性更高。

（三）心得

隨著網絡安全威脅的日益嚴峻，傳統的紅藍對抗模式逐漸轉向紅藍合作，這正是Purple Team的核心理念。Purple Team不僅僅是將紅隊（Red Team，進攻方）與藍隊（Blue Team，防守方）組織在一起，而是創建一種更加協作和透明的工作方法，使雙方能夠在持續的測試和回饋中實現共同進步。

首先，Purple Team強調的“合作”而非“對抗”具有深刻的戰略意義。傳統的紅藍對抗測試往往會引發團隊之間的資源封鎖，紅隊不願意透露入侵手法，藍隊則不願透露檢測策略，這導致了訊息孤島的形成。而Purple Team的受信代理人（Trusted Agent）模型，則讓部分藍隊成員直接參與到紅隊的行動中，這不僅增強了透明度，還加快了藍隊的威脅檢測速度。這一策略不僅讓組織能夠實時修補漏洞，還能增強紅隊和藍隊之間的信任和協作關係。

其次，Purple Team的連續攻擊模擬（Continuous Attack Simulation）方法有效提升了企業的檢測和回應能力。許多安全漏洞並不是一次性發現的，而是由於系統的控制漂移（Control Drift）而重新出現的。而Purple Team會持續進行攻擊模擬，藉助像Atomic Red Team這樣的工具來檢查系統的持續性防

禦，確保漏洞不會因系統變更而復發。這種持續的測試方法，可以幫助企業實現更高效的安全補救和回應，大大縮短了平均檢測時間（MTTD）和平均回應時間（MTTR）。

從心理和文化的角度來看，Purple Team的成功依賴於信任和透明的企業文化。在傳統的對抗性測試中，紅隊和藍隊之間經常存在對立情緒，而Purple Team的協作模式強調共享訊息，雙方必須擁有共同的目標和願景。例如，紅隊需要詳細記錄所有的行動過程，並提供時間戳和操作日誌，以便藍隊能夠回顧和檢查檢測的效果，這不僅增強了訊息的可追溯性，也能幫助企業在事後審查和改進策略時更有依據。

然而，Purple Team的實施也面臨一些挑戰。資源分配和信任的建立是推動這一策略的兩個關鍵因素。由於紅藍雙方需要共享數據和方法，如果沒有信任作為前提，藍隊可能會擔心紅隊會“作弊”，紅隊則擔心藍隊提前設置陷阱（如IP封鎖、異常檢測觸發）。為了解決這一問題，許多企業引入Zero Knowledge Attacker模式，紅隊必須假設自己不知道內部的基礎設施，這使得藍隊的檢測變得更加真實且具有挑戰性。

總結而言，Purple Team的出現標誌著網絡安全方法從「對抗性」轉向「協作性」，這不僅提高了組織的威脅檢測和回應速度，還幫助企業構建更加彈性和動態的安全體系。隨著網絡威脅的日益複雜，企業不僅要關注技術上的變革，還要在文化上實現信任、協作和透明的轉型。這不僅能夠降低企業的運營成本，還能加快創新步伐，讓企業在威脅不斷增長的網絡環境中保持競爭優勢。

五、Securing LLMs: Spear Phishing the Machines

（一）講者

Mohit Sharma（Chief Consultant， Atea AS）

（二）內容摘要

在AI與LLM快速崛起的背景下，相關的技術開發與應用逐漸成為產業趨勢，並對企業的運作模式與資安風險管理產生深遠影響。隨著OpenAI、Azure AI、AWS等雲端服務供應商的技術成熟，企業如今可以更容易地訪問和使用LLM，無論是構建AI解決方案、生成內容，還是作為自動化支援系統的一部分，但這也帶來了新的資安挑戰，特別是在訓練數據的質量、LLM的操作風險、資料隱私與存取控制等方面，必須制定更嚴謹的保護措施。

在AI模型的建構過程中，訓練數據的選擇和管理是首要關注點。數據中毒（Data Poisoning）成為攻擊AI模型的常見手段，對此，開發者需要確保數據的來源清晰可追溯，並對數據進行標註和審查，以降低模型被惡意操控的風險。數據標註和標籤的準確性將直接影響LLM的決策能力和生成內容的質量。此

外，知識產權的問題亦不容忽視，尤其是當企業依賴外部平台（如OpenAI或雲端服務）時，如何保障AI生成的內容與結果的知識產權，已成為企業在簽訂合約時需要考量的重要議題。

在LLM的操作層面，企業通常會部署一個多層架構來保護這些模型的運行安全性。這些架構包括模型託管API、編排層（Orchestration Layer）和數據層（Data Layer）。具體來說，模型託管API負責接收用戶請求，將這些請求路由到編排層，而編排層則作為企業的業務邏輯控制中心，決定LLM的訪問權限和操作範圍。例如，常見的工具如LangChain和Llama Index會在該層中發揮作用，控制LLM能訪問哪些數據、能執行哪些行動。數據層則包括向量數據庫（VectorDBs）和數據倉儲（如Elastic或Databricks），存儲著LLM需要訪問的關鍵業務數據。

由於LLM的架構和傳統的應用系統架構有所不同，LLM的數據平面與控制平面之間的界限不如傳統應用那麼明確。在傳統的應用中，控制平面（Control Plane）和數據平面（Data Plane）通常是獨立的，但在LLM應用中，控制邏輯和數據處理常常是交織的，這導致控制命令和數據流的分離更加困難。例如，在SQL注入攻擊中，攻擊者試圖在數據中插入特殊字符（如單引號、AND和分號等），以逃逸出原始語境，進而操縱後端控制系統的行為。這一現象在跨網站腳本攻擊（XSS）和跨站請求偽造（CSRF）中也經常出現，而類似的問題也可能在LLM的Tokenization和嵌入Embedding過程中出現，攻擊者可能試圖操控LLM生成的預測結果。

由於LLM會計算語意距離來預測後續單詞，這一點常被用作Prompt Injection的突破點。攻擊者可能通過在提示語中注入關鍵字或命令，讓LLM生成不正當的回覆或操作。這類攻擊的後果可能是數據洩露、誤導性回應甚至生成惡意代碼。為了對抗這種攻擊，企業應在Tokenization過程中增加檢查機制，確保不被可疑的提示劫持。

（三）心得

LLM的興起代表了一種嶄新的AI應用方式，其影響範圍已擴展到業務自動化、用戶體驗改善和內容生成等多個領域。隨著LLM的應用範圍不斷擴大，安全風險和控制挑戰也與日俱增。企業需要在訓練數據的治理、API的存取控制和業務編排的權限管理方面進行全方位的檢討，並考慮將AI威脅建模與資安風險控制結合，形成一套完善的防護架構。通過對LLM的控制平面和數據平面的分離、提示注入攻擊的檢測、身份驗證的強化以及風險評估的細化，企業才能有效控制LLM的資安風險，保障數位資產的安全性。

心得及建議

隨著人工智能（AI）技術的迅速發展，其在資安領域的應用日益廣泛，特別是在威脅檢測、異常行為分析、惡意軟體檢測和自動化回應等方面。AI能協助安全運營中心（SOC）透過大規模Log數據分析，在早期發現威脅並主動防禦。然而，AI本身的風險與挑戰也逐漸浮現，例如AI模型的數據偏見、對抗性攻擊（Adversarial Attack）和模型可解釋性問題。此外，生成式 AI（如 ChatGPT）也被駭客用於自動生成釣魚郵件和惡意代碼，進一步增加了資安威脅的複雜性。如何有效利用AI並同時防範其風險，已成為當前資安領域的核心議題。

2024年12月5日筆者參加於陽明交大舉辦的「2024 LLM產學技術論壇」，會中國網中心宣布即將推出LLM的線上服務，而陽明交大則成立了「myLLM 產學聯盟」，這一系列舉措彰顯出當前AI LLM技術的產學合作熱潮。不僅是科技界，百工百業正積極導入AI技術，以尋求創新和競爭優勢。這場論壇傳遞的明確信號是，AI技術的應用已成為不可逆的趨勢，未來LLM技術的發展與應用將對各行各業產生深遠影響。

從資安角度出發，AI時代的來臨勢必帶來新的挑戰與風險，特別是與LLM技術的應用與保護息息相關，回顧聯徵中心的發展歷程，自2019年開始自建SIEM平台，致力於收集與監控Log資料，這為後續的資安基礎打下了堅實的地基。隨後，聯徵中心逐步導入EDR（端點檢測與回應）等多項資安產品，形成一套完整的資安防護架構。2023年10月，聯徵中心正式啟用MDA戰情室（Monitoring and Detection Analytics），此舉不僅大幅提升了資安威脅的即時感知能力，更標誌著資安作業進入聯合作戰的全新階段。

然而，面對AI與LLM技術的快速發展，僅僅依靠被動監控與應變已不足以應對未來的資安挑戰。未來，聯徵中心的資安策略將從自動化向智能化轉型，這不僅需要持續優化SIEM和MDA的自動化作業，更需要導入AI技術，例如智能威脅偵測、異常行為分析和自動化應變，使系統具備更高的自我學習與適應能力。為了實現這一目標，培育具備AI和LLM應用專長的人才至關重要。人才的成長不僅僅是技術上的精進，還需強化AI威脅建模、生成式AI的資安風險識別、LLM提示注入攻擊的防範等技能。透過內部的專案推進和產學合作，讓資安團隊具備前瞻視野與專業技能，以迎接未來的AI資安時代。

總結來說，LLM技術的崛起既是機遇也是挑戰。聯徵中心將持續深耕SIEM和MDA戰情室的自動化作業，同時推動智能化資安解決方案的研究與實踐，並培育AI資安人才，以全方位防禦的策略迎戰AI資安新挑戰，為數位資產提供更堅實的保護屏障。