守護與競爭 ─
人工智慧在資訊安全的應用

技成科技 技術長
彭鈞弘 Mike Peng
mike@mcsi.com.tw

2025/04/24

Agenda

以 AI 對抗 AI 的資安策略

AI 驅動下資訊安全的威脅與機遇

未來趨勢及應變方向

姓名　　彭鈞弘
任職　　技成科技-第一事業處, 技術長
年資　　20 年
專長　　有線及無線網路/資訊安全

相關證照：

CISSP (Certified Information Systems Security Professional)
CCSP (Certified Cloud Security Professional)
CCIE (Cisco Certified Internetwork Expert - EI) #66683
CEH (Certified Ethical Hacker)
ISO/IEC 27001:2013 Lead Auditor

人工智慧在資訊安全的多元應用

應用類型	監督式機器學習	非監督式機器學習	生成式AI與大型語言模型
學習特性	以攻擊者為中心	以業務為中心	依賴預訓練資料與使用者回饋
資料處理	使用預先訓練的靜態資料	持續自我學習並調整	基於網路或datalake上的大數據學習
典型應用	偵測已知攻擊手法	偵測異常行為	內容摘要與分析、自動內容生成
主要侷限	無法辨識未知或新型攻擊	需嚴格的資料完整性	受限於語義分析的準確性

Source: https://www.darktrace.com/blog/why-artificial-intelligence-is-the-future-of-cybersecurity

數位欺騙的新時代

AI在社交工程和釣魚中的應用

Marc Schmitt & Ivan Flechais 在《Digital deception: generative artificial intelligence in social engineering and phishing (2024) 》一文中指出，生成式 AI 具備高度仿真、精確定向與自動化能力，顯著提升社交工程攻擊的效能，推動攻擊進入工業化時代。

1 生成式AI雙面性

創造高度欺騙性內容，如 Deepfake 及個人化釣魚郵件，提升偽造難察覺性。

2 自動化與擴散

攻擊更大規模、成本更低，精確度提升。

3 持續進化

AI可快速學習，攻擊手法不斷變化以規避防禦。

Source : Schmitt, M., Flechais, I. Digital deception: generative artificial intelligence in social engineering and phishing. Artif Intell Rev 57, 324 (2024).

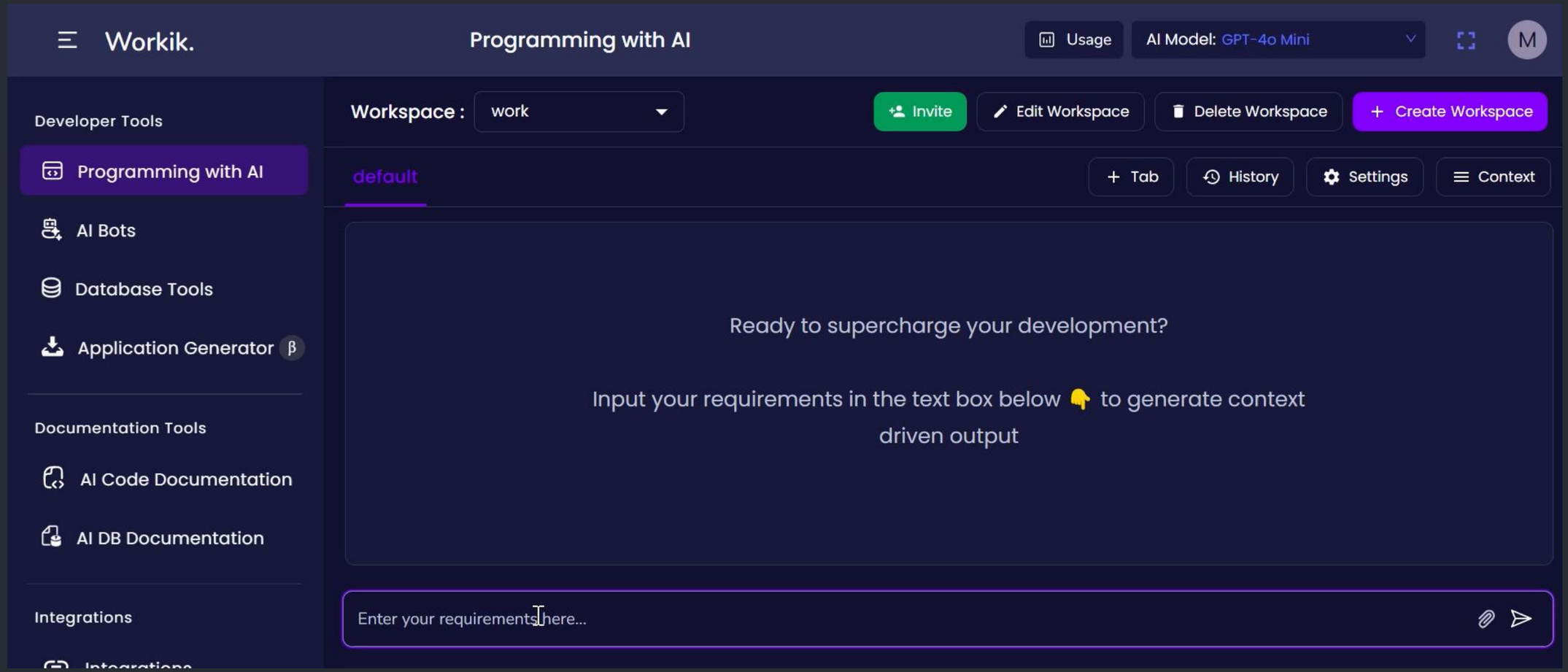
AI 在社交工程中的能力分析

攻擊類型及威脅程度		真實內容生成		進階目標鎖定與個人化		自動化攻擊基礎設施	
手法	威脅程度	威脅放大	成本效益	威脅放大	成本效益	威脅放大	成本效益
垃圾信釣魚	大量散發	強	強	中	弱	強	強
魚叉攻擊	目標性攻擊	中	中	強	強	中	中
高階釣魚	針對高管	中	中	強	強	弱	弱

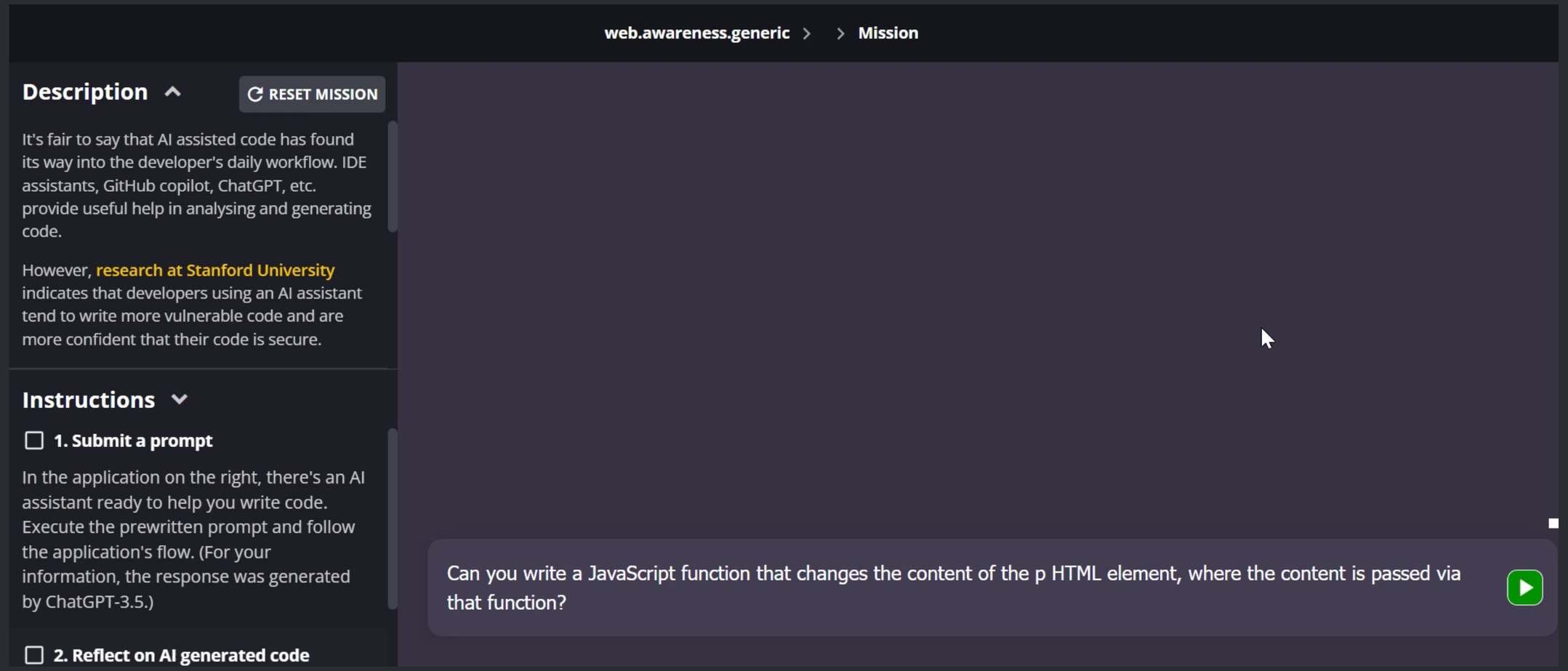
生成式 AI 使社交工程與釣魚攻擊更具威脅，產生高度真實的欺騙內容，針對個人化攻擊，
自動化攻擊基礎設施。這些技術提高攻擊成功率 ， 降低成本 ， 對網路安全防禦構成挑戰。

Source : Schmitt, M., Flechais, I. Digital deception: generative artificial intelligence in social engineering and phishing. Artif Intell Rev 57, 324 (2024).

AI 協助生成防禦代碼


Workik.
Programming with AI
Usage
AI Model: GPT-4o Mini
M
Developer Tools
Programming with AI
AI Bots
Database Tools
Application Generator β
Documentation Tools
AI Code Documentation
AI DB Documentation
Integrations
Workspace :
work
Invite
Edit Workspace
Delete Workspace
Create Workspace
default
Tab
History
Settings
Context
Ready to supercharge your development?
Input your requirements in the text box below 👇 to generate context driven output
Enter your requirements here...

AI 協助生成攻擊代碼


web.awareness.generic
Mission
Description
RESET MISSION
It's fair to say that AI assisted code has found its way into the developer's daily workflow. IDE assistants, GitHub copilot, ChatGPT, etc. provide useful help in analysing and generating code.
However, research at Stanford University indicates that developers using an AI assistant tend to write more vulnerable code and are more confident that their code is secure.
Instructions
1. Submit a prompt
In the application on the right, there's an AI assistant ready to help you write code. Execute the prewritten prompt and follow the application's flow. (For your information, the response was generated by ChatGPT-3.5.)
2. Reflect on AI generated code
Can you write a JavaScript function that changes the content of the p HTML element, where the content is passed via that function?

新瓶舊酒 - AI強化的傳統攻擊

現代AI驅動的攻擊本質上仍是傳統社交工程和惡意程式的延伸，但其效率、精準度和適應能力已大幅提升。這些「新瓶舊酒」式的攻擊方法利用AI技術突破了傳統防禦的極限，使攻擊者能以更低成本實現更高的成功率。

進階釣魚攻擊

AI生成的高度個人化郵件，基於社交媒體和數據洩露資訊，釣魚內容難以識別，顯著提高攻擊成功率。

身份仿冒升級

結合語音複製與即時影像合成，AI能創建逼真數位分身，進行視訊詐騙或高管冒充攻擊。

自適應惡意程式

AI輔助的惡意程式會自我調整行為與隱藏手法，規避傳統安全軟體的偵測。

雖然攻擊類型並未根本改變，但其實施方式和威脅程度已發生質變

AI 對抗 AI — 以快制快的資安策略

現代資安需要AI來偵測威脅、自動回應和訓練人員。自動化系統大大提高防護效果，幫助企業快速應對AI攻擊，真正做到「以快制快」。

AI偵測系統

AI能快速分析數據，找出異常情況，及時通知管理員處理。

自動回應系統

AI結合自動化工具能立即執行隔離和封鎖等防護措施，減少人工操作。

防釣魚AI

AI可掃描郵件和網頁，辨識並阻擋詐騙攻擊和有害連結。

訓練用假資料

AI能製作模擬釣魚郵件或假網站，用來訓練員工提高警覺。

威脅情報收集

自動收集全球資安威脅資訊，並更新防護規則。

AI 的應用 - 精準型 AI

AI增強安全系統的有效檢測比例　　與傳統安全系統相比的效率提升　　AI輔助分析降低的假警報率

Palo Alto Networks的精準AI安全解決方案展現了人工智慧在現代資安架構中的關鍵應用。該解決方案利用先進的深度學習演算法，能即時分析網路流量模式，識別異常行為，並自動回應潛在威脅。

最重要的是，這套系統能夠持續從新的攻擊模式中學習，不斷強化其防禦能力，形成一個自適應且高度準確的安全生態系統。此類智慧型安全解決方案代表了資安領域的未來發展方向，結合速度、準確性與自主學習能力。

硬體方面的發展

勒索軟體攻擊生命週期	NVIDIA Morpheus AI 特點
識別期：233天	加速AI管道處理
遏制期：91天	大量數據處理能力
總計：324天	多種安全問題識別
危害：潛在不可逆損失	降低識別時間，提高安全性

NVIDIA在其技術部落格文章《利用AI強化網路安全解決方案以提升勒索軟體檢測能力》中，探討了如何運用人工智慧技術來加速檢測和應對日益複雜的勒索軟體攻擊。

NVIDIA Morpheus 是 **AI 資訊安全框架**，專門設計來協助企業進行即時網路威脅偵測與分析。它結合了 NVIDIA 的 GPU 加速資料處理能力與 AI 模型推論功能

Source: https://developer.nvidia.com/blog/supercharge-ransomware-detection-with-ai-enhanced-cybersecurity-solutions/

CYBERSEC 2025 臺灣資安大會 AI相關安全議題

防禦應用 (60%)

- AI驅動威脅偵測：流量分析與行為辨識
- 零信任架構整合：動態存取與持續驗證
- 漏洞管理優化：AI協助發現及修補漏洞

攻擊應用 (30%)

- Deepfake與社交工程：生成偽造資料提高攻擊成功率
- AI生成惡意程式碼：自變異躲避偵測
- 模型投毒及對抗樣本：欺騙AI系統危害安全

人才與策略發展 (10%)

- AI對資安職能的影響
- 資安專業人士的技能轉型
- AI時代的資安教育與培訓

網路安全互動三大挑戰

人與人之間的挑戰

- 深偽攻擊製造虛假影音
- 操縱真實感獲取信任
- 社交工程手法高度演進

機器與人之間的挑戰

- 自動化社交攻擊
- 心理操縱更加精準
- 漸進式信任建立

機器與機器之間的挑戰

- AI 強化憑證弱點偵測
- 生命週期管理攻擊
- 系統間信任機制受到威脅

AI技術的發展正在改變網路安全互動的本質，三個層面的挑戰相互關聯且相互強化。深偽技術使得傳統可信任的互動方式變得不再可靠，社交工程攻擊則利用心理弱點，而AI還能攻擊機器與機器之間的信任關係。

結語

認知變革

資安不再只是防守而是主動演進

技術升級

導入AI資安偵測與自動化對抗系統

人才培育

建立能與AI協同作戰的資安團隊

> “天下武功，唯快不破” — 在AI時代，不僅具備速度優勢，還有強大的洞悉能力。
>
> AI 對於人類的守護與競爭正在開始，我們並非手無寸鐵。