生成式AI風潮下關鍵資安議題

蔣建軍

前言

生成式人工智慧（Generative AI）因其卓越的文本生成、影像合成及語言處理能力，迅速成為全球關注的技術焦點。從聊天機器人到深偽影像生成，生成式AI的應用覆蓋了社交媒體、商業、醫療、娛樂等多個領域。然而，隨著技術的普及，其潛在風險也日益突顯。在生成式AI的發展過程中，數據隱私洩露、假訊息泛濫、網路安全威脅升級等資安問題已成為各界關注的焦點。例如，深偽技術被廣泛應用於創建偽造視頻，用於詐騙或政治操控；AI生成的惡意軟件攻擊降低了網路防禦的門檻，對傳統資安工具形成巨大挑戰。因此，探討生成式AI風潮下的資安議題，對於技術的健康發展與應用規範化至關重要。

生成式AI的崛起與應用現狀

（一）生成式AI的技術基礎

1. 深度學習（Deep Learning）

深度學習是生成式AI的核心技術之一，其主要特點是構建多層神經網路來模擬人腦的學習機制。透過大量數據的訓練，深度學習模型可以學習複雜的模式與規則，並運用於不同場景下的創造性生成。例如，GPT模型通過預測文本的下一個詞，生成連貫且具有上下文意義的段落。這種技術突破使生成式AI在自然語言處理（NLP）、影像合成及語音生成等領域展現了強大的能力。

2. 生成對抗網路（GANs）

生成對抗網路是一種利用生成器和判別器相互競爭的框架。生成器負責創建假數據，而判別器則評估數據真假，並將結果反饋給生成器進行優化。GANs的主要應用包括生成高度擬真的圖像、修復損壞的數據以及增強現有圖像的質量。這種技術已被廣泛應用於電影製作、虛擬現實與醫學圖像分析，為創意產業和科研領域帶來了全新可能性。

3. 自然語言處理（NLP）

NLP是一門讓機器理解、處理和生成人類語言的技術。透過結合語法分析、語義理解與上下文建模，NLP使生成式AI能夠創建極具人性化的對話系統、自動翻譯工具及文本摘要應用。如今，NLP技術已廣泛應用於智慧客服、語音助理和知識圖譜構建，並為用戶提供更流暢且個性化的互動體驗。

（二）生成式AI的應用場景

1. 商業領域

在商業領域，生成式AI的應用日益多元化。企業利用AI生成產品描述、廣告文案和社交媒體內容，顯著降低人力成本並提升市場行銷效率。此外，AI還能幫助企業分析消費者行為，透過精準的大數據分析協助決策，優化產品設計與營銷策略。例如，零售業中使用AI預測季節性需求波動，大幅提高供應鏈的敏捷性與可靠性。

2. 醫療與科學研究

在醫療領域，生成式AI已經開始重塑診斷與治療的方式。例如，通過生成病理圖像模擬，AI能幫助醫生更快地發現早期疾病跡象，提升診斷準確性。此外，在藥物開發中，生成式AI利用化學結構模擬加速新藥的研發進程，降低實驗成本。對於科學研究，生成式AI還可以自動化數據處理流程，幫助科學家快速提取有價值的洞見。

3. 娛樂與創意產業

在娛樂產業中，生成式AI已經成為動畫製作與電影特效的重要工具。AI能自動生成高質量的虛擬場景與角色，縮短製作週期並降低成本。同時，AI生成的音樂和劇本為創意產業提供了豐富靈感。此外，遊戲設計中運用AI進行場景生成與角色塑造，不僅提升了遊戲的沉浸感，還滿足了玩家對個性化體驗的需求。

生成式AI帶來的關鍵資安議題

（一）數據隱私洩露

1. 數據蒐集的透明性與合法性

訓練生成式AI模型需要大量用戶數據，包括文本、圖片甚至音頻，而這些數據可能包含用戶的敏感資訊。若企業在數據收集過程中未充分告知用途或未經用戶同意，將可能違反數據隱私法規（如GDPR）。此外，數據供應鏈的不透明性會導致部分數據來源不明，進一步提高洩露風險。例如，有些公司可能從第三方購買含有個人信息的數據，造成潛在的法律與道德問題。

2. 逆向推斷攻擊

當前生成式AI模型存在被逆向推斷的風險，攻擊者可以通過分析模型的輸出結果，推斷其訓練數據的敏感信息。例如，在與ChatGPT的互動過程中，若輸入數據包含私人信息，攻擊者可利用多次測試獲得與該信息相關的片段，進而推斷完整資訊。這樣的漏洞可能被利用於惡意活動，威脅用戶隱私和數據安全。

（二）假訊息與深偽技術濫用

1. 假訊息傳播的規模與速度

AI生成的假訊息擁有極高的擬真性，使受眾難以辨別真假，尤其是在社交媒體平台上，假訊息的傳播速度往往超過真實信息。這對政治、經濟與公共安全構成重大威脅。例如，在重大選舉期間，AI生成的假新聞可能影響選民意向，甚至引發社會分裂。此外，這些假訊息常利用人們的情感共鳴進行擴散，加大控制假訊息的難度。

2. 深偽技術的道德挑戰

深偽技術能模擬人物的聲音與外貌，創建高度擬真的影像或視頻。這類技術若被濫用，可能製造虛假證據，用於勒索、抹黑或操縱輿論。例如，攻擊者可能生成某公眾人物的不雅視頻，試圖毀壞其形象；或者偽造企業高管的聲音進行詐騙交易。這些行為不僅侵犯了個人隱私，也對信任體系和法律運作提出了嚴峻挑戰。

（三）生成式AI驅動的網路攻擊升級

1. 針對性網路釣魚攻擊

傳統的網路釣魚郵件往往因語法錯誤或內容缺乏針對性而被識破。然而，生成式AI能根據特定目標生成高度定制化的釣魚信息，例如模仿受害者熟悉的語氣和格式發送詐騙郵件。這種技術降低了受害者的警惕性，極大提高了釣魚攻擊的成功率，尤其是在企業環境中可能導致機密數據洩露。

2. 自動化漏洞挖掘與利用

生成式AI可用於自動生成惡意代碼或尋找軟件系統的潛在漏洞。例如，攻擊者可能訓練AI模型專門分析開源代碼庫，尋找未修補的安全漏洞並生成可利用的惡意代碼。這種行為不僅加速了攻擊準備過程，也對企業現有的資安防禦機制構成了新的挑戰。

（四）技術濫用與法律挑戰

1. AI生成內容的責任歸屬

當生成式AI創建的內容對社會或個人造成損害時，責任應由誰承擔？目前法律尚未對此給出明確指引。例如，若AI生成的虛假信息引發大規模社會混亂，開發者是否應該承擔部分

責任？或是使用該技術的用戶需全權負責？這類問題涉及技術倫理與法律層面的深層考量，並需要各國共同制定清晰的規範來解決。

2. 跨境監管困難

生成式AI應用的全球性特徵，使其在法律層面面臨跨境監管的難題。例如，一個在甲國開發的AI技術可能被乙國的惡意用戶濫用，而甲國法律對此可能無法追責。再者，各國對於隱私與內容審查的法律標準不一致，導致技術濫用的灰色地帶擴大，增加全球協同治理的難度。

生成式AI資安挑戰的應對策略

（一）加強技術層面的防禦機制

1. 模型訓練的數據匿名化與加密處理

為了減少數據洩露的風險，AI模型訓練過程應採用數據匿名化技術，確保個人信息在數據集中無法被識別。此外，數據的存儲與傳輸應採用高標準的加密技術，如端對端加密與零信任架構，降低攻擊者入侵數據供應鏈的可能性。同時，應建立數據審查機制，定期檢查訓練數據是否包含敏感或非法來源的內容，以確保合法性與安全性。

2. 設計防禦性AI對抗攻擊

為了應對生成式AI驅動的自動化攻擊，可以設計專門的防禦性AI系統，用於主動檢測並阻止潛在威脅。例如，利用生成對抗網路（GANs）對抗深偽內容，通過判別器提升對虛假數據的識別準確性。同時，企業應部署基於AI的威脅情報平台，實時監控網路環境中的異常行為，及早發現並遏制攻擊行動。

（二）加強法規與政策的制定與執行

1. 建立生成式AI內容標識與追溯機制

為了抑制假訊息與深偽內容的泛濫，各國政府應推動生成式AI內容標識的法規。例如，要求所有AI生成的圖像、音頻和視頻需嵌入無法移除的數字水印，以標明其為AI創作。此舉有助於提高公眾對生成內容的辨識能力，並加強對惡意內容的追溯能力。此外，需制定對應的法律框架，對濫用技術的行為進行處罰，震懾潛在的不法分子。

2. 促進國際合作與規範統一

由於生成式AI應用具有全球性，單一國家難以獨立應對其資安挑戰。因此，各國應通過多邊協議合作，建立統一的技術與法律標準，例如共同制定數據隱私保護與內容審查的規範。國際組織如聯合國或G20可扮演協調角色，推動各國在技術治理和執法層面的合作，應對跨境濫用問題。

（三）提升公眾與企業的數位素養

1. 強化公眾辨識假訊息的能力

為了減少假訊息的傳播，需提升公眾對生成式AI內容的辨識能力。政府與教育機構可開展專項培訓，普及AI技術的基本知識，並教授如何檢測深偽內容。例如，介紹利用數位工具檢測圖片與影片真偽的方法，或者分析網絡訊息來源的可信度。此外，透過廣泛的社會宣導活動，讓民眾對生成式AI的可能風險有更深刻的認識，增強防範意識。

2. 推動企業加強資安培訓與投資

企業應認識到生成式AI帶來的雙面效應，將資安作為其技術應用與發展的核心策略之一。例如，定期舉辦內部資安培訓，幫助員工了解針對性網絡釣魚攻擊與其他生成式AI驅動的威脅。同時，企業需加大資安技術的投入，例如部署先進的威脅檢測與回應（EDR）系統，並建立完善的應急響應計劃，確保在面臨攻擊時能迅速恢復業務運營。

（四）促進生成式AI技術的倫理與透明性

1. 建立AI開發的倫理指導原則

技術開發者應遵循「負責任AI」的理念，在設計與應用生成式AI時考慮其可能引發的倫理問題。例如，確保模型訓練過程中的數據來源合規，並避免使用可能侵害隱私或違反人權的數據。此外，應建立多方參與的倫理審查委員會，定期評估AI技術的潛在風險與社會影響，並提出改進建議。這不僅有助於增強公眾對技術的信任，也可減少不當使用的可能性。

2. 提升技術透明性與可解釋性

生成式AI技術的高複雜性常使其成為「黑箱」，公眾難以了解其運作機制。為了解決此問題，開發者需加強技術的可解釋性，例如通過提供簡化模型的運算邏輯或發布相關技術文檔，讓用戶能夠了解AI做出決策的基礎。同時，應開展更多的技術透明化實踐，例如舉行公開演示或論壇，促進公眾對生成式AI技術的理解與監督。

結語

生成式AI的迅速崛起雖然為各領域帶來了創新與變革，但其潛在的資安風險不容忽視。從數據隱私洩露、假訊息泛濫，到網路攻擊升級與法律挑戰，生成式AI的風潮既是技術的里程碑，也對資安治理提出了全新挑戰。面對這些威脅，我們需要採取多層次的應對策略，包括技術強化、政策完善、教育推廣與倫理規範，確保生成式AI的發展能真正造福人類社會。