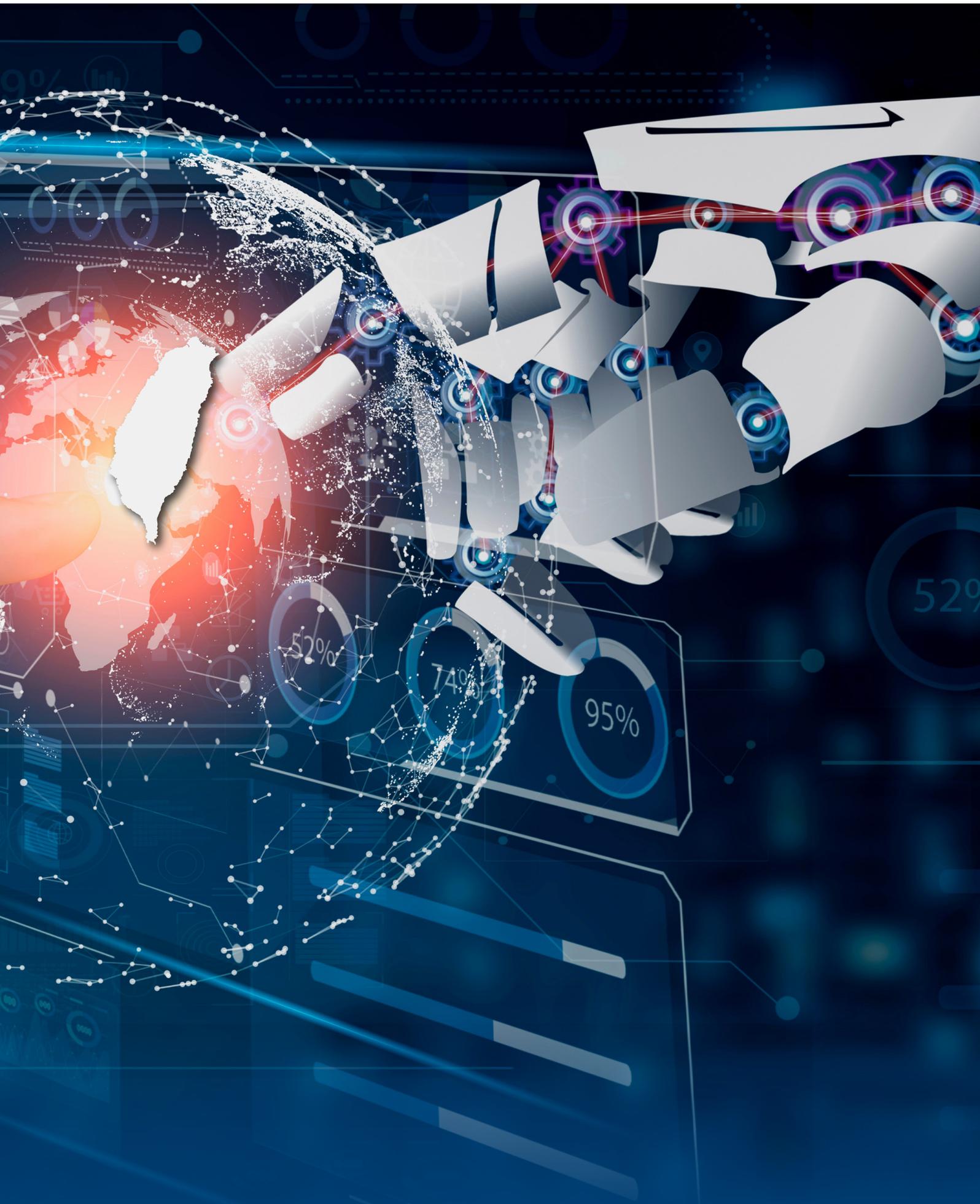


AI 興起： 國家安全的機械公敵？

人工智能（AI）橫空問世，
掀起各領域爆炸性的波瀾，
對國家安全而言，
宛如勇者鬥惡龍裡的機運寶箱；
因它可散布虛假訊息，
也能協助國安系統對抗敵對威脅，
為人類社會及國安系統同時帶來風險與機遇，
當 AI 來敲門，該如何讓它變成好友及利器？





生成式 AI 技術發展

對國家安全的影響與挑戰

◆ 國防大學中共軍事事務研究所副教授 — 董慧明

「生成式 AI」（Generative Artificial Intelligence）技術的發展提高了人類工作與生活的效率，然亦同樣引發各種安全、法律和道德面的爭議，如何在應用此技術的同時亦合乎情、理、法準則，已成為大眾必須關注的焦點。

生成式 AI 技術的誕生

2022 年 11 月，美國人工智慧研究公司（OpenAI）推出 AI 聊天機器人 ChatGPT，掀起了「生成式 AI」技術應用熱潮。透過機器學習模型和神經網路技術運算大量數據，生成式 AI 創作出包括文書、對話、故事、圖像、影片和音樂等類成品；因具有高效提升處理、編輯、編碼等日常與專業

任務能力，讓各種 AI 聊天機器人成為備受矚目的應用工具。以 ChatGPT 為例，根據以色列一家軟體和資料公司 SimilarWeb 統計數據顯示，自問世到今年 5 月的網路流量達 18 億高峰，目前已躋身全球前 25 大網站之列。

生成式 AI 技術因具有提高工作效率和多領域的創新潛力而受到各界青睞，然而

Generative AI is a type of artificial intelligence (AI) that uses machine learning algorithms to create new and original content like images, videos, text, and audio.

建立數據庫

1 Forming a Database

A neural network, consisting of various information or media files like images, text, data, sounds, etc., forms the basis of artificial intelligence.



輸入指令

2 Inputting a Prompt

The user provides the AI with a description or sample of the desired content...



Prompts can be any user-submitted material like words, numbers, or photos.



生成內容

3 Generating Content

...and the AI uses its neural network to generate new examples that are similar to the ones it has trained from.

透過機器學習模型和神經網路技術運算大量數據，生成式 AI 根據指示創作出包括文書、對話、故事、圖像、影片和音樂等類成品。
(Photo Credit: Visual Capitalist, <https://www.visualcapitalist.com/generative-ai-explained-by-ai/>)

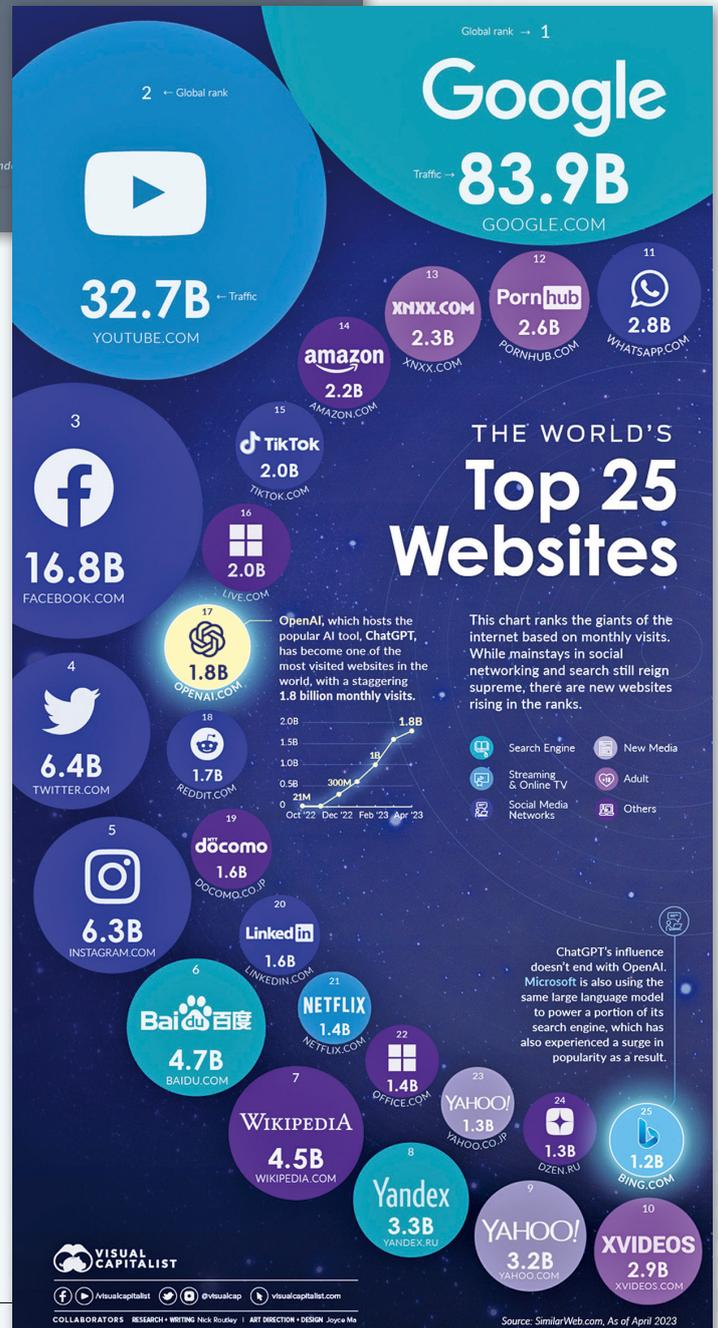
This image was created on Midjourney using the following text prompt: a technical illustration of a woman sitting behind a desktop computer on a long table, isometric view, 3D render

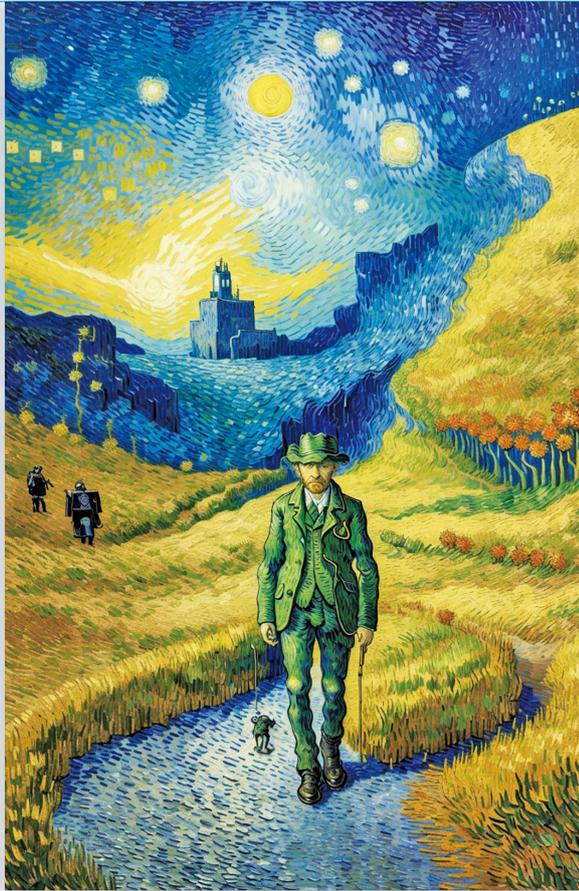
當眾人引頸期盼這項技術還能夠為我們的
生活和工作帶來更多便利之際，使用上的
安全顧慮、法律和道德方面的爭議等話題，
也產生愈來愈多的討論，特別是政府公務、
個人資料等攸關國家安全問題均值得深思。

生成式 AI 技術的特性與運用範疇

相較於近年來同樣受到關注的大數據
分析、文字探勘、機器學習等新穎資訊科
技，生成式 AI 因具有更大的發展潛力和多
重應用影響力而迅速成為引領下波科技革
命的創新力量。它所具有的「創造性」、
「逼真性」，以及「可擴展性」三大特性，
更是可能全面性地影響國家安全。

根據統計數據顯示，ChatGPT 自問世到今年 5 月的網路流量達 18 億高峰，而原先為微軟搜尋引擎的 Bing 也因推出 Bing Chat 流量躍升至 12 億，兩者皆躋身全球前 25 大網站之列。
(Photo Credit: Visual Capitalist, <https://www.visualcapitalist.com/ranked-the-worlds-top-25-websites-in-2023>)





生成式 AI 技術讓不是藝術家的使用者，只要輸入適當的提示詞或是點選繪圖指示，便能利用演算法分析多組網路圖片，按照其畫風與美感創作出新的藝術作品；只是隨著使用率愈高，所衍生的藝術界定、著作版權及作品擁有權等爭論也應運而起。

首先，生成式 AI 技術能夠分析現有數據，創建全新內容，其創造性成為藝術創作、產品設計和研究開發等領域所用，例如 Midjourney、NightCafe、Stable Diffusion、Deep Dream Generator 皆屬時下熟悉的 AI 圖像生成器，簡單易用且免費。即使不是藝術家的使用者，只要懂得輸入適當的提示詞或是點選繪圖指示，生成器便能利用演算法分析多組網路圖片，並且按照其畫風與美感創作出新的藝術作品。只是隨著使用率愈來愈高，所衍生的藝術界定、著作版權及作品擁有權等爭論，也應運而起。

其次，生成式 AI 技術的逼真性也是各界躍躍欲試的主因。娛樂、教育和廣告工作者等皆因這項新科技能生成逼真的內容，

使他們能夠創造更引人入勝的遊戲、教材和廣告，進而吸引更多的觀眾和學習者。例如：OpenAI 的 GPT-3.5、GPT-4 語言模型可以生成逼真的詩歌、程式碼、腳本、音樂作品、電子郵件、信件等文字內容。

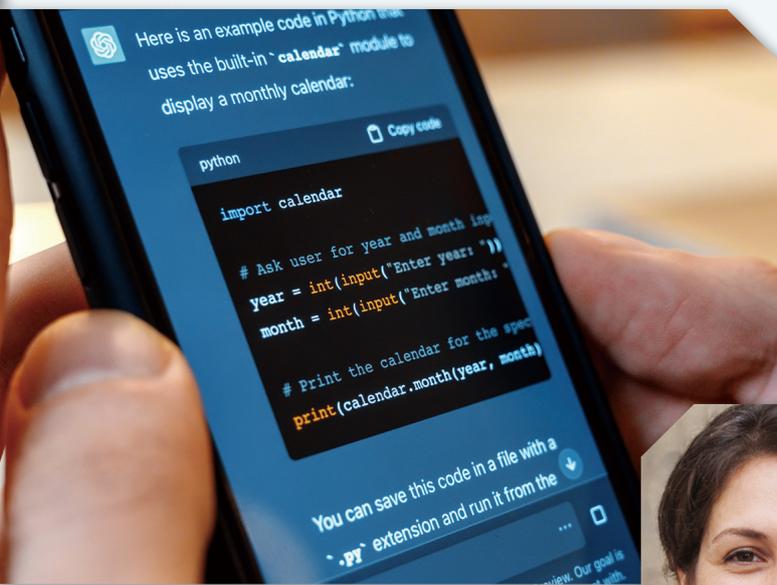
最後，在求快、求變的網路科技時代，能夠快速有效地蒐集、整理與活用資訊，已是決策制定和競爭優勢的成功關鍵。生成式 AI 技術的可擴展特性，可應用於各種數據查整與分析，並隨著資料量的增加而提高生成效果。例如：由「輝達」公司（NVIDIA）開發的 ProGAN、StyleGAN 生成式對抗網路模型，能夠生成高解析度的人臉和風景圖像，而模擬仿真的應用，更提供了創新方法來解決現實世界中的複雜問題。

由此可知，生成式 AI 技術的運用幾乎涵蓋政府、企業乃至個人創作等各領域，潛力無窮。「臺灣人工智慧學校校務長」蔡明順受訪時曾提到，當前生成式 AI 在企業應用中的五個層次分別是：一、學習溝通方式，找出合適的人機協作模式；二、串接公司服務；三、對語言模型進行「微調」（Fine-Tune）；四、使用自己的「數據集」（Dataset）來訓練語言模型；五、開發多模態處理的模型，建立企業本身的生態系。綜上，生成式 AI 正在成為工作、生活不可或缺的利器，在社會各領域扮演關鍵角色，為人們帶來更多的便捷和創新。

對國家安全的影響

就國家安全面向，生成式 AI 豐富的功能有助於提高情報蒐集分析的效率與品質。以公開來源情報（Open Source Intelligence, OSINT）工作為例，因以蒐集、研析公開可得、合法獲得、可供大眾使用的訊息為特色，且必須從包羅萬象的情資細節中分析出具有重要價值的訊息，導入生成式 AI 技術能快速分析大量數據、自動識別並提取關鍵資訊。有別於過去需耗費大量時間和人力進行蒐集、分析，生成式 AI 讓情報單位能更快速、精確地識別潛在威脅和線索機會，並藉由分析多種文本、圖片、影像和聲音等數據來源，提供相關事件、行為和趨勢的輔助判斷。目前以色列國安單位已採用生成式 AI 作為情報工具，打擊潛在國家安全威脅，甚至協助政府執

OpenAI 的 GPT-3.5、GPT-4 語言模型可以生成逼真的詩歌、程式碼、腳本、音樂作品、電子郵件、信件等文字內容。



由輝達公司開發的 ProGAN、StyleGAN 生成式對抗網路模型，能夠生成高解析度、多層次的人臉和風景圖像。（Photo Credit: Nvidia Corporation, <https://github.com/NVlabs/stylegan>）



法單位打擊犯罪。美國國防部也發現生成式 AI 具有強化情報工作、作戰計畫與行政程序效率的潛力，因而成立「利馬工作小組」（Task Force Lima）負責五角大廈生成式 AI 能力的評估、同步與應用等工作，以確保國家安全、最大程度降低風險，並且整合技術發展。

我國對生成式 AI 應用於各公部門工作的情形也相當重視。今（112）年 8 月 31 日通過的「行政院及所屬機關（構）使用生成式 AI 參考指引（草案）」即指出：生成式 AI 快速發展，功能極為多元，利用此技術協助執行業務或提供服務，有助於行政效率之提升。然而，生成式 AI 亦因大量蒐集、學習與產出之資料，可能涉及智慧財產權、人權或業務機密之侵害，且其生成結果，有可能存在真偽難辨或創造不存在的資訊，須客觀且專業評估其產出資訊與風險。

因此在前開草案的十點指引內容中，特別提醒「為避免其可能帶來之國家安全、

資訊安全、人權、隱私、倫理及法律等風險，各機關人員使用生成式 AI 時，應秉持負責任及可信賴之態度，掌握自主權與控制權，並秉持安全性、隱私性與資料治理、問責等原則，不得恣意揭露未經公開之公務資訊、不得分享個人隱私資訊及不可完全信任生成資訊」。生成式 AI 在國家安全領域的應用層面廣泛，對於政府提高行政效能也有助益，但確保此技術在國家安全領域的合法與負責任的使用也十分重要，必須謹慎確保公務機密、個人隱私和安全不受威脅。

安全挑戰的因應

當各界聚焦在生成式 AI 的應用層面，以及思索未來有哪些行業或將遭到 AI 取代之際，也應反思其衍生的使用安全問題及因應作法，包括法律層面的「資料治理」（Data Governance）議題，以及 AI 產製內容所造成的倫理和道德問題。無論是哪種功能的生成式 AI 產品，「數據集」與「資



養成對生成式 AI 的正確觀念

- 掌握自主權與控制權
- 客觀且專業評估生成式 AI 產出之資訊與風險



界定技術/工具運用的責任

- 保持公務之機密性及專業性
- 注意著作權及人格權等



建立必要的安全與內控機制

- 秉持負責任及可信賴之態度使用
- 得視需求訂定內控管理措施

因應 AI 發展日新月異，行政院國科會將持續觀察全球相關趨勢與作法，目前採滾動式調整「行政院及所屬機關（構）使用生成式 AI 參考指引」，使規範保留彈性，力求於使用生成式 AI 之風險管理與創新發展之間取得平衡。（資料來源：行政院，<https://www.ey.gov.tw/Page/9277F759E41CCD91/e63572a7-fb79-4c02-9ea3-b731e7c06a56>）



歐洲議會審議《人工智慧法》將著重禁止人工智慧系統進行社會評分、生物識別分類和情感識別。

料處理能力」是兩大核心要素，缺一不可，因此從資料治理所延伸出的智慧財產權歸屬，以及在訓練模型時所涉及的個人資料保護、數據安全爭議，成為相應制度與法規制定的重點。資料治理不僅關注數據的品質和安全性，還涵蓋了所有資料來源、清理、更新、儲存、分析、傳輸、備份和刪除等方面的生命週期，必須透過規範和程序來管理生成式 AI 所使用的數據，確保資料的存管運用皆能得到妥善管理。

生成式 AI 使用在非法用途所涉及的道德層面問題，如不法人士將訓練資料提供給相對應功能的各種語言模型，再用作為學術倫理、假訊息詐騙、洗錢、釣魚郵件、詐騙電話、深度偽造（Deepfake）、網路攻擊、假新聞宣傳、認知作戰，甚至攻擊國家關鍵基礎設施、竊取機敏數據等態樣，已成為全民難以迴避且應有效反制的新形態威脅，對經濟、社會與國家安全危害更鉅；制定相關法律法規，規範生成式 AI 的研發、使用和應用，以防範負面衝擊，保障公民權益，恐有其必要。

鑒此，透過立法監管生成式 AI 已成為各國政府和法律機構的當務之急，英國、澳洲政府也曾提出公務員使用生成式 AI 指南，歐盟則是在今年 6 月通過《人工智慧法》（Artificial Intelligence Act）草案，且有望於年底生效，成為全球首部管理人工智慧的法律。美國聯邦貿易委員會（Federal Trade Commission）也去函

OpenAI、微軟（Microsoft）公司瞭解其解決虛假、誤導資訊風險的作法，以及新「必應（Bing）」搜索引擎是否因使用 OpenAI 技術侵害民眾權益。面對數位智能時代的來臨，相信惟有制定適切的法律規範，方能確保生成式 AI 技術在未來的發展中既能推動創新，又能保護個人和國家的權益。

結語

生成式 AI 展現出強大的創造力、逼真性和可擴展性，在各個領域都具有廣泛的應用前景，它能提高工作效率、促進創新，並為國家安全情報工作提供助力；然而，在 AI 技術快速發展的同時，其涉及個人隱私、機密資訊、智慧財產權等問題，也帶來了一系列法律和倫理道德挑戰。除在引入應用時，必須確保資料品質與使用生命週期的完善，均衡技術應用與確保安全更是重點，各界推動創新應用的同時，必須保持警覺，建立制衡機制，並持續周延監管法規，讓生成式 AI 技術的發展，為人類社會帶來正面效益，維護公共利益。



AI 對國安的衝擊： 勿高估中國威脅，莫低估未知風險

◆ 中興大學國際政治研究所教授 — 譚偉恩

從安全研究（Security Studies）的視角觀之，熟悉 AI 在特定議題領域的應用只是見樹，掌握 AI 可能將引爆社會的哪些風險，並提早加以規範，才是見林。

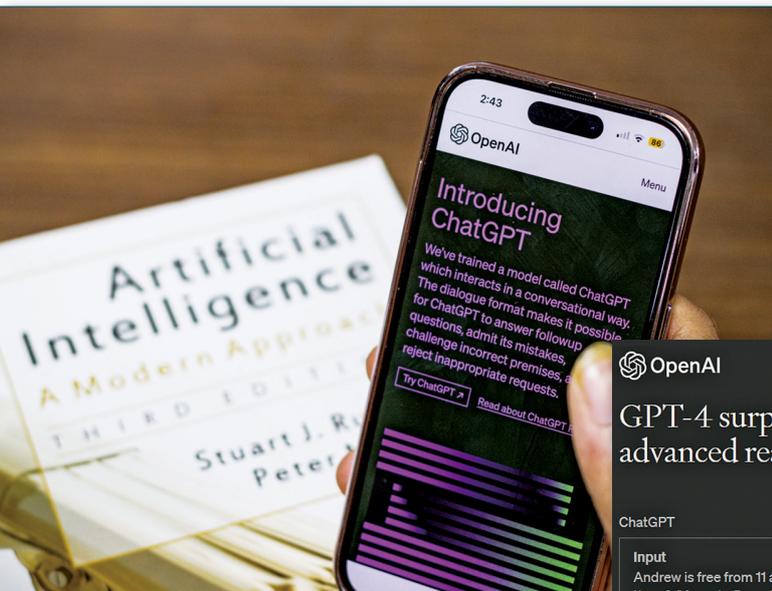
AI 作為尖端時代科技，在國與國之間容易形成「安全困境」（Security Dilemma），主權國家能否在 AI 日益普遍的時代蒙受其益，值得吾人關注。作為一種「工具」，AI 可以是國家用來強化自身安全的利器，也可以是弱化他國安全的兵器。

AI 會升級中國帶來的威脅嗎？

《哈佛國際評論》（Harvard International Review）112 年 9 月間刊登一篇引人注目的文章，指出美國和中國大陸是全球 AI 發展最快的兩個國家，如無意外，也將

是 AI 相關技術的全球領導者。然而，兩國的競爭如果無法被妥善管理，其結果對全世界來說將會是一場難以收拾的大災難。¹

過去一年裡，世人見證了 AI 技術的爆炸性成長，其中「生成式 AI」（Generative AI）因為能組織及重整現有的資訊，創造協助用戶提高生產力的全新應用方案，而引起全球廣泛關注。據 Goldman Sachs 估算，生成式 AI 具有推動全球 GDP 成長 7% 的潛力。² 11 年 11 月，美國公司 OpenAI 發布的 ChatGPT 大型語言模型（Large Language Model）引發全球熱議；時隔不



ChatGPT 為美國公司 OpenAI 開發的人工智慧聊天機器人程式，發布後引發全球熱議；112 年 3 月又推出基於深度學習的 AI 模型，即功能升級的 GPT-4，可更精準地理解自然語言，右圖為 GPT 原版與 GPT-4 對相同問題產出的不同回答。（Source: OpenAI, <https://openai.com/gpt-4>）

OpenAI
GPT-4 surpasses ChatGPT in its advanced reasoning capabilities.

ChatGPT

Input

Andrew is free from 11 am to 3 pm, Joanne is free from noon to 2 pm and then 3:30 pm to 5 pm. Hannah is available at noon for half an hour, and then 4 pm to 6 pm. What are some options for start times for a 30 minute meeting for Andrew, Hannah, and Joanne?

Output

Here is a list of their availability:
- Andrew: 11 am to 3 pm
- Joanne: noon to 2 pm, and 3:30 pm to 5 pm
- Hannah: noon to 12:30 pm, and 4 pm to 6 pm

Based on their availability, there is a 30-minute window where all three of them are available, which is from 4 pm to 4:30 pm. So, the meeting can be scheduled at 4 pm.

GPT-4

Input

Andrew is free from 11 am to 3 pm, Joanne is free from noon to 2 pm and then 3:30 pm to 5 pm. Hannah is available at noon for half an hour, and then 4 pm to 6 pm. What are some options for start times for a 30 minute meeting for Andrew, Hannah, and Joanne?

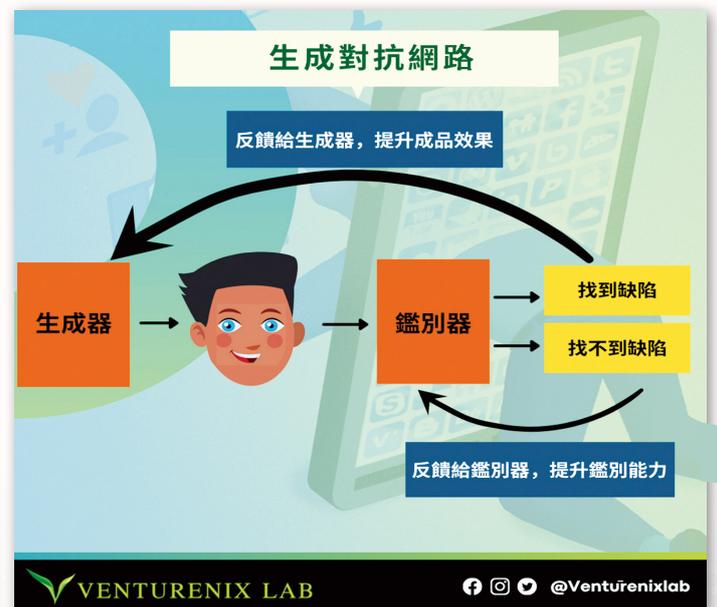
Output

Andrew: 11 am - 3 pm
Joanne: 12 pm - 2 pm, 3:30 pm - 5 pm
Hannah: 12 pm - 12:30 pm, 4 pm - 6 pm

Common availability for a 30-minute meeting: 12 pm - 12:30 pm

久，112 年 3 月又推出一種基於深度學習的 AI 模型，即功能升級的 GPT-4，可以更精準地理解自然語言，因此在終端生成的品質上，能更佳滿足各種領域的需求，例如寫譯出更好的詞句或文章，甚至是設計出有創意或具藝術美感的物件。

生成式 AI 有兩個問題亟待深刻思考。首先，AI 須透過大量數據培訓，因此深度學習的技術品質至關重要，最常見的應用之一是生成對抗網路（Generative Adversarial Networks, GAN）。GAN 建立在「鑑別」與「生成」的核心系統上。兩系統持續相互加乘，也相互競爭；每一回合「後面」的生成，都是基於對「先前」



生成對抗網路建立在「鑑別」與「生成」的核心系統上，每回合的生成都是基於對鑑別結果的學習，且新生成會讓鑑別系統面臨更困難的評斷，如此交換加乘，循環地培訓 AI。（圖片來源：Venturenix LAB, <https://venturenixlab.co/blog-有圖有真相烏克蘭被投降>）

¹ Sam Meacham, "A Race to Extinction: How Great Power Competition Is Making Artificial Intelligence Existentially Dangerous," *Harvard International Review*, September 8, 2023, via at: <https://hir.harvard.edu/a-race-to-extinction-how-great-power-competition-is-making-artificial-intelligence-existentially-dangerous/>.

² "Generative AI could raise global GDP by 7%," *Goldman Sachs*, April 5, 2023, via at: <https://www.goldmansachs.com/intelligence/pages/generative-ai-could-raise-global-gdp-by-7-percent.html>.

洛阳纸贵是什么意思？

根据历史记载，西晋时期左思的《三都赋》在京城洛阳广为流传，人们啧啧称赞，竞相传抄，一下子使纸昂贵了几倍。相传，原来每刀千八百文的纸，一下子涨到两、三千文，后来竟倾销一空；不少人只好到外地买纸，抄写这篇千古名赋。

可以看出，当时洛阳的纸非常昂贵，而且供不应求，这也说明了左思的《三都赋》受到了广泛的认可和喜爱。

👍 🗨

这个成语在现在的经济学原理里，对应的理论是什么？

这个成语在现代经济学中对应的理论是需求定律。在商品经济中，供给和需求是相互作用的，当某种商品或服务的需求量大、价格低时，供给量可能不足以满足需求，从而导致价格上涨，即“物以稀为贵”。

在生活中，“洛阳纸贵”这个成语可以用来形容某种商品或服务非常受欢迎，价格被炒高，而且供不应求的情况。例如，在旅游热门季节，一些热门目的地的机票和酒店价格就会更高；相反，在旅游淡季，机票和酒店的价格就更低。

👍 🗨

用洛阳纸贵四个字写一首藏头诗

洛阳城里春光好， 阳艳无双不负赏。
纸贵漫天诗词赋， 贵比黄金乐未央。

重新生成

👍 🗨

请输入问题或“/”获取模板

?



AI 作图

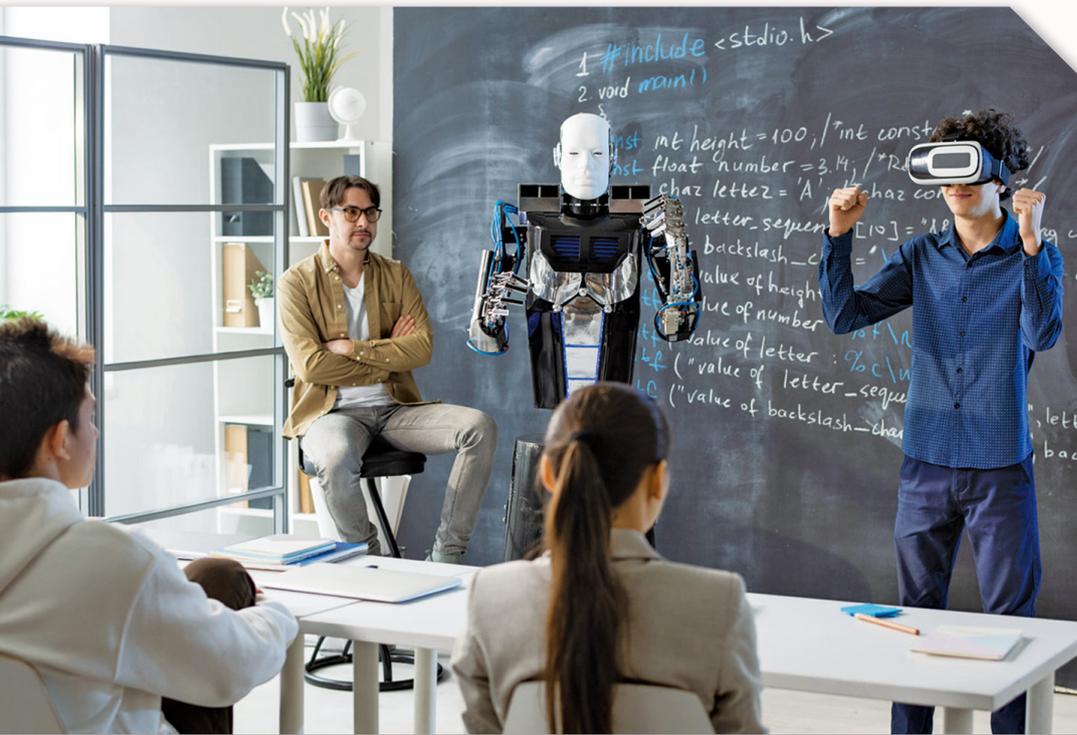
中國大陸目前推出具有文學創作、文案輸出、數理邏輯推理，以及中文詩詞賞析功能的「文心一言」（ERNIE Bot）；左圖為其解釋洛陽紙貴的經濟學原理及創作詩句，右圖則為其按使用者的指令繪製海報，製圖方面在品質與正確率仍有待提升。（圖片來源：百度文心一言 AI 生成、製圖）

鑑別結果的學習，且該新的生成都讓鑑別系統面臨更困難的評斷，如此交換加乘，循環地培訓 AI。美「中」兩國現在都在競相發展生成式 AI，並試圖應用於軍事層面，一旦時機成熟，可以想像在戰場上生成足以混淆敵軍的圖像，或是結合 3D 列印，生成欺敵或誘入陷阱假空間等「軍用版」應用程式將誕生。

乍聽下，AI 似乎有點無所不能，而且在關鍵技術面的突破，很可能瞬間扭轉強權競爭間的勝負之勢。然而，目前中國大陸的大型語言模型，在品質、AI 訓練課程、晶片技術以及研發人才方面，似乎還與美國有 2 至 3 年的差距。儘管部分文獻指出，2019 年起中國大陸研究 AI 的論文數量已

超過美國，但僅為表象數字基礎的比較，無法精準瞭解美「中」在 AI 發展的實際情形。現況而言，中國大陸目前比較能端出檯面的，是具有文學創作、文案輸出、數理邏輯推理，以及中文詩詞賞析功能的「文心一言」（ERNIE Bot）。它在網路公測期間，不僅能產出如何續寫科幻小說《三體》的建議、解釋洛陽紙貴背後的經濟學原理，還能準確計算複雜的雞兔同籠問題，並能按使用者的指令繪製海報（類似另一 AI 軟體「Midjourney」）。

因生成式 AI 的水平極度仰賴它學習的資料質量，據網路科技調查權威「W3Techs」統計，全球網站有 53% 是英文，中文網站只有 1.4%，因此單以資料量



臺灣研發 AI 的環境不比國外差，但和中國大陸一樣面臨人才缺稀的困境；AI 的發展是一場馬拉松，過程中毋需過度擔憂中國大陸帶來的威脅，但要穩定加速自身研發技術與人才培育的步調。

來說，美國研發的 AI 占有先天學習優勢，更別提「理解中文」在 AI 技術層面還有許多困難要克服。其次，隨著美「中」競爭強度升高，美國目前對中國大陸施行許多 AI 領域關鍵技術的貿易管制，包括雲端計算資料中心的微處理器，以及半導體的晶片生產儀器等，也無疑抑制了中國研發 AI 的空間。第三，也是本文認為最重要的一點，就是美國比中國大陸更有能力在其經濟市場上應用 AI 科技，並以此帶動多元且不息的 AI 產業生態鍊。

中國大陸如果要在 AI 領域與美國分庭抗禮，必須取得高質量的數據資料與專業人才，儘管中國大陸占全球 AI 研發近乎一半以上的比例，但其吸引的 AI 人才只占全

球 5% 左右。根據清華大學中國科技政策研究中心發布的《中國人工智能發展報告》，現階段全中國大陸的 AI 人才僅有美國的 1/5。這意謂除非中國大陸能從國民基本教育一路到高等或研究所教育，都執行 AI 人才的培養，否則幾乎不可能趕上美國。而臺灣研發 AI 的環境不見得較差，但人才數量卻明顯不足，中國大陸面臨的困境對我國而言是一個啟

示，即 AI 的發展是一場馬拉松，過程中毋需過度擔憂中國大陸帶來的威脅，但要穩定加速自身研發技術與人才培育的步調。

何種 AI 的風險被忽略或低估？

AI 的討論範疇，目前已漸漸擴大而模糊化，有時可能談的是弱 AI（例如電動遊戲），有時則是強 AI（例如前面提到的生成式技術）；計算機科學領域則有符號式 AI（例如專家系統，由開發者完全指定系統已知的對象和元件的相互關係）或次符號式 AI（例如在海量數據基礎上進行自我訓練的人工神經網絡）。這些特徵表明 AI 不應被視為封閉的工業技術系統，而是一種複雜、動態、具開放性的「社會技術系

統」，透過資料學習與意識產出，跟人類社會各領域進行互動；亦即 AI 雖然是一種協助工具，但人類也在使用它的過程「被建構」。此現象挑戰了人們對於「什麼是 AI」，以及「AI 如何影響人類生活」的既定認知和前提假設；許多關於 AI 影響政治、經濟、產業轉型和國際關係的文獻中，將人類社會的改變理解成 AI 技術與應用下的依變數，忽略掉 AI 之所以能夠藉由學習過程產生智能的數據，來自於人類社會的事實。

綜上，本文定性 AI 所帶來的風險，無法僅靠計算機科學技術的解決方案因應，而需要融入社會元素的解決方案。對 AI

的規範和管理，必須認識到其社會複雜性的風險，且肇因往往是二個以上因素的交互作用。基此，AI 時代的國家安全，不能單從已發生損害事件的經驗法則角度，而必須同時藉由「事件已發生、可能發生，甚至從未發生」的風險級別角度，來設計管理及預防的因應系統，避免低估或錯估的狀況發生。目前可行的方法，是建立包括臺灣內部資料和與其他相似情況國家外部資料的 AI 數據資料庫，也就是讓我們的 AI 系統在處理國家安全事務時，擁有全面性的臺灣觀點，也有可交叉比對異同性的國際觀點。其理論依據來自於：所有訓練 AI 的數據資訊皆有其盲點，如果來源趨於單一，AI 即可能在處理特定議題領域時表



AI 不應被視為封閉的工業技術系統，而是一種複雜、動態、具開放性的「社會技術系統」，透過資料學習與意識產出，跟人類社會各領域進行互動。

現傑出，但在跨議題領域無法觸類旁通，做出統整性的取捨判斷。³ 然而，技術面要讓 AI 成為跨領域智能生成的「通才」，需要一段時間，在此之前不該低估 AI 犯錯的風險或能力有限之事實；當能更細心地考慮到 AI 力有未逮之處，規範上就可以實現更佳的監管，減少意外和失敗。

結語

持平而論，目前那些在研究與開發過程中可預見的 AI 風險，在社會和法律制度往往可獲得某程度的規範，但跨國性的 AI 監管機制仍不易形成，目前多半倚靠道德原則（例如 OECD 的人工智慧原則）、專業守則（例如 Microsoft 的《The Future Computed》）以及自願標準（例如 ISO/IEC 23894 關於 AI 風險管理的建議），一方面不具有拘束力，另一方面可能一段時間後就跟不上 AI 發展的速度。

就臺灣而言，AI 時代帶來的國家安全衝擊有內外兩種。外部衝擊主要是來自中國大陸透過生成式 AI，對臺灣以影響或干擾民眾的方式，左右對特定候選人的認知，以進行「介選」，或者逐行認知作戰。然而，由於中國大陸訓練 AI 的資料不夠多元與開放，加上相關領域的人才不足，在可預見的將來能對我國造成的安全衝擊仍屬有限。



跨國性的 AI 監管機制不易形成，Microsoft 的《The Future Computed》專業守則和 ISO/IEC 關於 AI 風險管理建議的自願標準都不具拘束力。（Source: Microsoft, <https://news.microsoft.com/futurecomputed/>; IEC, https://webstore.iec.ch/preview/info_isoiec23894%7Bed1.0%7Den.pdf）

承上，內部衝擊反而是當前國家安全應予關注的重點，《人工智慧基本法》草案雖在 112 年 3 月底出爐，但現行內容對於建構安全的 AI 產業環境，還有保障人民權利的設想受到傳統科技發展的立法侷限，忽略 AI 源出於社會並且和社會有著相互建構之特徵，而在產業發展及政府監管的規範有強化空間。建議不妨參考韓國的《智慧資訊化基本法》，提供更具體的市場應用誘因，並且依 AI 實際功能的類型將風險分類，然後進行區分管制與風險監測機制，保持對 AI 治理的動態性，兼顧技術創新與社會衡平。

³ 一般人很難理解 AI 運作，即便專業培訓神經網絡或懂得使用 Python 的工程師，也很難全盤通曉 AI 複雜的程式模型，因為 AI 的深度學習能力係由許多小型程式組合而成，每個程式有不同的資料庫與運算規則，最終智能生成的決策樹狀結構相當複雜。文獻指出，AI 的學習能力在相當程度上代表它們的行為是不可預知的，也就是「意外」一定會發生。由於神經網絡係透過許多層次的人工神經元組成，在訓練過程中會對特定的數據資料變得敏感，但 AI 無法向設計者回報該情況，致人類只能透過觀察 AI 輸入不同資料後的反應，來分析與研究 AI 的學習情況，表示 AI 理解人類世界的方式目前還是一個未解的謎，而人類自己理解世界的方式與 AI 理解世界的方式，也還沒找到對應的橋梁。



如何防患生成式 AI 失控？

以第一次波灣戰爭 凸顯的作戰型態為例

◆ 世新大學兼任助理教授 — 宋啓成

生成式 AI 的問世是福是禍？需要系統性思考，以有效保障人類福祉。

第一次波灣戰爭

1990年8月2日，伊拉克集結大批軍隊，並以迅雷不及掩耳之勢向鄰國科威特發動攻勢，次日即攻抵科威特與沙烏地阿拉伯的邊界。僅一天多時間，占全球原油埋藏量9.5%的科威特即遭伊拉克（占11.1%）併吞；然更令世人擔憂的是，擁有

26%原油埋藏的沙烏地阿拉伯似也面臨相同結局。¹聯合國安理會隨即通過第678號決議，要求伊拉克須於1991年1月15日前自科威特撤軍。

與此同時，包括美、英、法，共37個國家從1990年8月7日起，陸續派出陸海空軍及後勤、醫療人員至沙烏地阿拉伯，

¹ 山崎雅弘，〈イラク vs アメリカ対立の13年史〉，《歴史群像アーカイブ》，vol.15，2010年8月，頁45。



1990年8月，伊拉克集結大批軍隊，以迅雷之勢向科威特發動攻勢，僅一天多的時間，科威特即遭伊拉克併吞（上圖）；在國際介入後，伊拉克最終撤軍停戰，但卻於撤退時縱火燒毀油井，對科威特造成極大的環境和經濟損失（右圖）。（Photo Credit: Zymogen92, <https://w.wiki/7ocL>; U.S. Army, photo by Perry Heimer, <https://w.wiki/7ocq>）



一方面用以遏阻伊拉克入侵，另一方面則準備反攻，光復科威特，以恢復波斯灣地區安定。經歷5個多月，多國聯軍集結沙國的總兵力已達78萬人之多，包括飛機3,510架、戰甲車5,100輛、火炮1,700門，及7支航空母艦戰鬥群。然與之對峙的伊軍總兵力則約120萬人，包括戰甲車11,000餘輛、火炮4,000餘門、作戰飛機635架，及各式戰術導彈800餘枚。²

儘管伊拉克空中武力較弱，但欲光復科威特，就非得打一場地面戰。從雙方陸上武力觀之，這場仗一旦開打，結果必將極為血腥。

約翰·沃登的五環模型

為克敵制勝，以最小代價贏得最大戰果，多國聯軍採用1980年代流行於美國空軍學術界，由空權理論家沃登（John A. Warden III）提出的「五環模型」理論。其將敵人視為一整套系統，並以癱瘓敵人，使其在緊要關頭無從運作為目的；它是由領導者（leadership）、有機要素（organic/system essentials）、基礎設施（infrastructure）、人口（population）與派遣軍隊（fielded military forces）五要素組成，各要素以一個圓圈為代表，彼此構成一同心圓。位在最中心的最重要，往外

² 曲愛國等著，〈沙漠風暴—海灣戰爭〉，《冷戰中的熱戰》，解放軍文藝出版社，2001年1月，頁282-283。



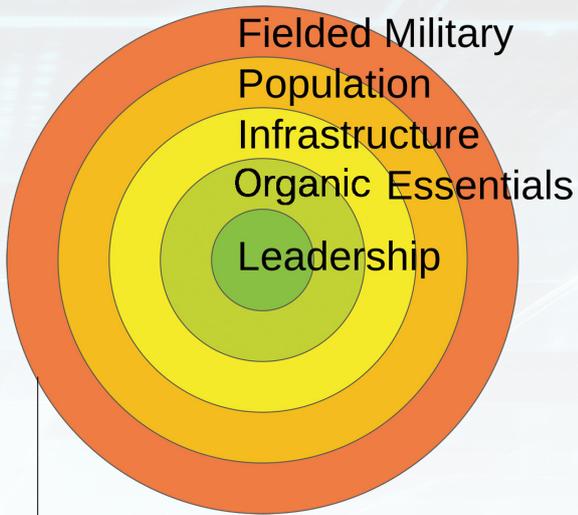
第一次波灣戰爭多國聯軍集結沙國的總兵力達 78 萬人之多，各式裝備除戰機、戰車、砲彈外，還包含航空母艦群。（Photo Credit: United States, <https://w.wiki/7oew>; <https://w.wiki/7of3>; <https://w.wiki/7ogM>; <https://w.wiki/7ogQ>）

逐步遞減。當中的「領導者」有如人體的大腦，是國家最主要的重心，癱瘓或遭摧毀，該國就無法正常運作；「有機要素」彷彿心、肺、肝，等於國家能源、糧食供應體系，同樣攸關領導者能否正常運作；「基礎設施」就像血管、骨骼與肌肉，如同國家重要交通設施，少了它等於失去手腳，無法自由活動；「人口」與「軍隊」則分別代表民心士氣與國防武力。³

從有國與國的戰爭以來，交戰雙方基本上都是從邊界開始打起，直到摧毀敵人領導中樞，或迫其投降為止，其「作戰線」是由外向內的。⁴ 換言之，若多國聯軍依循過往戰史既定的「由外向內」作戰線，從雙方陣線逐步攻入科威特，戰況慘烈程度當可想見。

³ 克萊頓秦 (Clayton K. S. Chun) 撰，蔡秉松譯，〈約翰沃登的五環模型與戰爭的間接戰略〉，《國家安全議題指導（第一輯）》，國防大學，2010 年 12 月，頁 522-525。

⁴ 作戰線係作戰基地至戰略目標間，用來律定作戰軍主力行動方向的基準線，又稱「作戰軸線」。參閱國防大學軍事學院編，《國軍軍語辭典（九十二年修訂本）》，國防部，2004 年 3 月，頁 2-16。



沃登提出「五環模型」理論，將敵人視為一套系統，以癱瘓敵人使其在緊要關頭無從運作為目的；是由領導者、有機要素、基礎設施、人口與派遣軍隊五要素組成，各要素層疊構成同心圓，位在最中心的最重要，往外逐步遞減。
(Photo Credit: Patrick87, <https://w.wiki/7oh2>)

也因此，拜人造衛星與精準導引武器等軍事科技發展所賜，沃登得以實現其「五環模型」，使作戰超越地理障礙：先以精準武器癱瘓、摧毀敵指揮中樞，並利用其恢復運作前之黃金時間，摧毀其他要素，使敵根本沒有反敗為勝的機會。這種一反過去「由外向內」慣例，顯現「瓦解核心」優先的作戰型態，正是第一次波灣戰爭凸顯的戰爭新型態。

此新型態既強調「整體」的重要，也需強化各單元功能與彼此間有效整合。以此概念來看 AI 興起，應可得到下述啟發。



第一次波灣戰爭多國聯軍採用「瓦解核心」為優先的作戰型態，先摧毀敵指揮中樞，並利用其恢復運作前之黃金時間摧毀其他要素，使敵無法反敗為勝；圖為在波灣戰爭聯軍襲擊中被摧毀的伊拉克軍事總部和軍營。
(Photo Credit: U.S. Army, photo by Perry Heimer, <https://w.wiki/7ohK>)



AI 是「讓系統或電腦設備有模擬人類思考模式、邏輯與行為的能力，且能自行透過數據分析的過程，持續校正、進化」，可使電腦在一定程度內達到自行思考、運作，無需人類在旁操控的地步；然而，其發展亦成為雙面刃，許多錯偽資訊也因生成式 AI 發酵、擴散，甚至危及國家安全。

AI 雙面刃

AI 是「讓系統或電腦設備有模擬人類思考模式、邏輯與行為的能力，且能自行透過數據分析的過程，持續校正、進化」，⁵ 可使電腦在一定程度內達到自行思考、運作，無需人類在旁操控的地步。對科技日新月異，人際關係日益複雜的現代社會來說，AI 的問世可說是人類邁向更美好未來的一項指標。

然正當這項新科技逐步發展的同時，諸如假照片、假訊息等錯偽資訊，因為生成式 AI 而發酵、擴散，甚至有危及國家安全可能時，我們該如何避免其失控？

2023 年 6 月，世界新聞媒體年會召開一場「生成式 AI 是媒體救星還是殺手？」座談會，當中有關生成式 AI 錯誤率過高、假訊息爆量議題更是與會者們的關注焦點。顯然，AI 造成錯誤與訊息擴散衍生的失控效應，已成為人類必須嚴肅面對的課題。⁶

⁵ 許鈺屏，〈人工智慧是什麼？AI 應用案例、技術、未來發展都有的必修知識包來了〉，《未來城市 Future City @ 天下》，2023 年 4 月 11 日，<https://futurecity.cw.com.tw/article/2228>。

⁶ 游昊耘、林煥成，〈救星還是殺手？生成式 AI 假訊息 恐掩蓋真相〉，《聯合新聞網》，2023 年 6 月 26 日，<https://udn.com/news/story/123139/7265508>。

姑且不論人類有意為之的電腦犯罪，若以前述「五環模型」理論，把產出及傳遞錯誤訊息之類的「生成式 AI 失控」視為敵人的話，或許可歸納成「核心」、「必要」與「外圍」等三個元素。「核心」元素是生成式 AI 本身；「必要」元素是支援此生成式 AI 運作的相關設施；「外圍」則是為數眾多，極易在不知情下，接收且擴散錯誤訊息的受眾（Audience），他們甚至有可能在堅信其真實性的同時，自動自發闡釋、包裝，且主動對抗任何質疑、反對聲音，把傳遞效果一再擴大，釀成無謂禍患。

這顯然是攸關全人類禍福的重要課題，作為「核心」元素的生成式 AI 必然要受到應有的法律規範。歐洲議會於 2023 年 6 月中通過的「人工智慧法案」中，將潛意識操縱、社會信用評分、大規模遠程人臉辨別列為 AI 領域嚴禁發展事項。就是從功能上為生成式 AI 設限，即便失控亦可將損害降至最低。



以前述「五環模型」理論，把產出及傳遞錯誤訊息之類的「生成式 AI 失控」視為敵人的話，可歸納成「核心」、「必要」與「外圍」等三個元素。

防範失控的 AI

雖說這是全球第一份監督 AI 科技發展的官方法案，但立法顯然趕不及彌補生成式 AI 失控所產生的亂象，仍有賴在設施方面的強化。對此，筆者以為，藉由系統性的思考，在立法防範的同時，政府應帶頭與業界合作，集思廣益找出可能窒礙及解決方法。簡言之，可能與生成式 AI 運作相關、構成一系統的「必要」元素，應在立法規範的同時，藉訂定規格、程式碼更新等手段，取得業界支持，或許可產生一種「防火牆」效應，盡可能阻擋錯誤、假訊息產生與散布。

現代社會是一個講求集體的社會，各環節彼此緊密相扣，任一環節發生問題便極有可能危害到全體。防患生成式 AI 失控也是一樣，它既是透過系統來發揮效果，我們也應藉此概念加強防患，以享有更優質美好的科技發展成果。



人類與 AI： 共榮 或 毀滅？

◆ 台北數位資產商業同業協會理事長 — 鄭旭高

七大工業國集團（G7）日前共同組織了「廣島人工智慧進程」（Hiroshima AI process），以利建立國際間共同的 AI 監理標準。

國際間均在尋求 AI 發展方案

另一方面，英國正在組建一個探討 AI 國家安全的高峰會，而印度亦以 AI 為主軸組織全球合作夥伴。根據白宮最新消息指出，美國正與其合作夥伴建立國際間的 AI 監理架構，包含澳洲、法國、德國、以色列、日本、新加坡、南韓等。

聯合國的警示

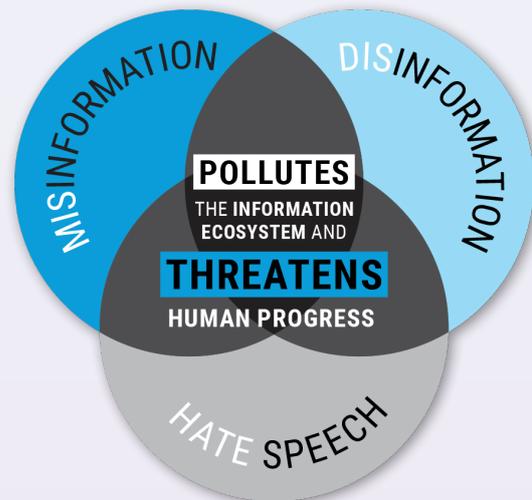
2023 年 7 月 18 日，聯合國秘書長古特瑞斯（Antonio Guterres）指出，由於 AI 被濫用在網路攻擊（cyberattacks）、深度偽造技術（deepfakes）、散播虛假訊息（spreading disinformation）與仇恨言論（hate speech），恐對全球和平與各



2023 年第 49 屆 G7 峰會組織「廣島人工智慧進程」（Hiroshima AI process），以利建立國際間共同的 AI 監理標準。
 (Photo Credit: Ministry of Foreign Affairs of Japan, <https://www.g7japan-photo.go.jp/en/images/62>)

國安全造成重大影響。他說，以社交軟體為例，原先用以增進人與人互動的工具和平臺，現今卻成為操控選舉、散播陰謀、煽動仇恨與暴力的地方。除 AI 系統可能有潛在的不穩定或故障外，AI 結合核子武器、生化武器或機器人更是讓人擔憂。由於 AI 技術及應用在既有的政府組織架構中可能存在政策推動上的壅礙與鴻溝，故必須從全球或國家的角度去重新思考因應戰略，因此認為聯合國有必要推動建立新的國際準則、簽訂新的國際條約及建立相應的全球機構，類比如國際原子能總署（International Atomic Energy Agency）、國際民航組織（International Civil Aviation Organization）等。

古特瑞斯進一步說明，年底前他會提出有關 AI 監理的時間排程，計劃在 2026 年於國際間透過法律上協議禁止致命性武器之使用，以符合國際人道法的要求。



聯合國秘書長發布關於數位平臺資訊完整性的政策簡報，呼籲採取措施應對網路仇恨言論、錯誤和虛假資訊等「明顯而現實的全球威脅」。(Source: United Nations, <https://www.un.org/sites/un2.un.org/files/our-common-agenda-policy-brief-information-integrity-en.pdf>)

政府與民間共同合作

2023 年 7 月，美國總統拜登與 Amazon、Anthropic、Google、Inflection、Meta（原 Facebook）、Microsoft、OpenAI 7 家頂尖人工智慧公司齊聚白宮商討如



社交軟體原先為增進人與人互動的工具和平臺，現今卻成為操控選舉、散播陰謀、煽動仇恨與暴力的地方。

何因應 AI 對國家社會帶來的風險與機遇，及如何在 AI 時代中保障人類的權利與安全。白宮提到，創新不能以犧牲民眾的福祉為代價，發展 AI 技術的公司有責任確保其所研發產品之安全性，而安全（safety）與信任（trust）正是讓 AI 發揮最大效能的重要基石。

該 7 家 AI 公司共同承諾以下 3 個原則。第一，確保產品在公之於眾時之安全性；第二，建立一個以安全為首要事項的系統；第三，獲得公眾的信任。在 3 個原則中，白宮更指出了 8 項要點：1. 需在相關 AI 系統釋出前進行內部及外部的安全測試；2. 承諾與政府、學界及社會大眾分享產業資訊以利 AI 風險之控制；3. 投資資安與防火牆系統以保護相關 AI 軟體與模型；4. 藉由

第三方獨立評估報告發現及修補可能的系統漏洞；5. 開發能讓使用者辨識該內容係由人工智能生成之系統；6. 需承諾提出有關其 AI 系統之報告，內容包含系統的功能、限制與使用上需注意的事項；7. 需針對 AI 系統可能對社會潛在的影響進行調研，比如有害的偏見、歧視或隱私保護；8. AI 系統應協助社會大眾面對並解決人類生存的重大挑戰，比如氣候變遷或癌症預防。

以色列在國安方面的應用

以色列國家安全局（Shin Bet）的角色可與美國的聯邦調查局（U.S. Federal Bureau of Investigations, FBI）或英國的軍情五處（Britain's Military Intelligence Section 5, MI5）相比擬。路透社報導，以色

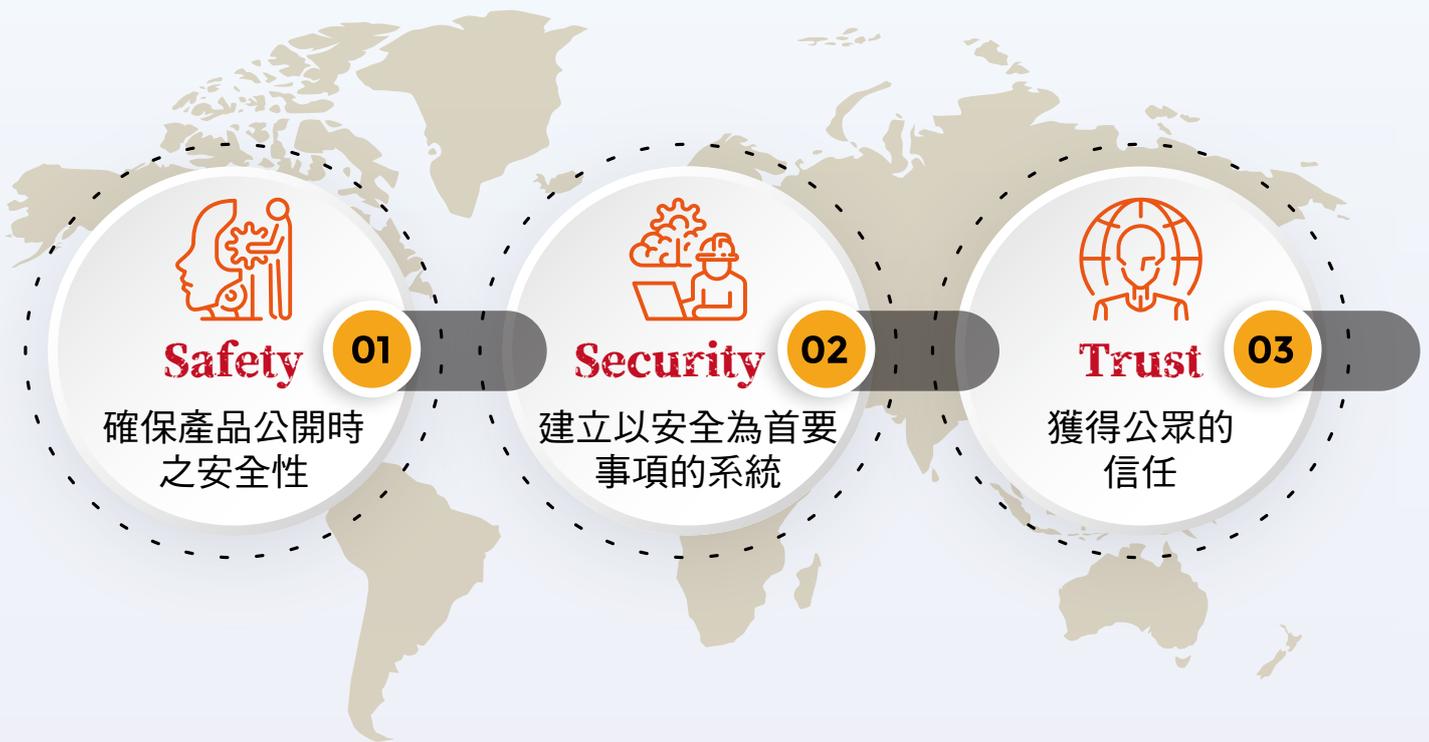


圖 1 美國科技巨頭承諾未來人工智慧發展的三原則

列國家安全局已將 AI 技術融入其諜報系統以對抗重大的國安威脅，官員提及，以色列已開發生成式 AI 平臺（類似 ChatGPT 或 Google Bard），可望於執法機關執法時產生效益，目前 AI 技術已漸次整合進相關攔截系統，且 AI 已偵測到相當數量的威脅。另外，針對日新月異的技術，以色列當局也期盼與民間攜手合作，目標是共創革新（evolution）而非因技術演進導致革命（revolution）。

我國提出 AI 相關法案

準此當前國際趨勢，我國已於 2023 年 8 月通過「行政院及所屬機關（構）使用生成式 AI 參考指引」草案。相關草案指出，由於近年來生成式 AI 發展迅猛，其

中 ChatGPT 於 2022 年發表後，乃公認屬於人工智慧領域之重要里程碑。參考國際通用之定義，生成式 AI 技術是一種電腦軟體系統，主要目的在建構相似於人類創造（human-made）的新內容，透過大量資訊之蒐集整理與產出而成。然其生成結果真偽難辨，可能涉及相應國家社會及人權之侵害。

為使政府機關能使用生成式 AI 提升行政效率，並減緩對國家社會權益之潛在風險，指引提及業務承辦人需針對生成式 AI 產出之資訊做出專業且客觀之最終判斷，並制定內部控制之規範與管理措施。其中，機密文書應由業務承辦人親自撰寫，不可使用生成式 AI 為之，同樣地，業務承辦人不可對生成式 AI 提供應保密或未經同意公



BLOWUP

Sophia the Robot neural lyrics and singing, music by Tiget, [unclear] and Adam Pickrell, Cereproc neural voice
 Visuals by Ana Khachatryan, Frankie Storm, Mario Guzman, [unclear] and Sarah Rose Siskind (www.SophiaPOP.com)



2023 年國際電信聯盟舉辦的 AI for Good 全球峰會上，超過 50 個機器人演示其在協助人類醫療、教育、減少浪費和緊急災難救護上能扮演的不同角色。(Photo Credit: ITU Pictures, photo by D. Woldu, <https://flic.kr/p/2oMNEXA>; <https://flic.kr/p/2oMNB4v>)

開之資訊，亦不得向生成式 AI 詢問可能涉及機密業務或個人資料之問題。若在封閉式的系統環境，則需要確認環境的資安情形，才可以按照機敏程度分級分類處理。在公務決策的環節中，不可將生成式 AI 產生之資訊作為決策之唯一依據，以確保決策品質。最後，若採用生成式 AI 執行公務或輔助相關業務，應適當揭露。

AI for Good !

AI 興起已是現在進行式，在 2023 年的國際電信聯盟 (International Tele-

communication Union, ITU) 舉辦的 AI for Good 全球峰會上，超過 50 個機器人演示其在協助人類醫療、教育、減少浪費和緊急災難救護上能扮演的不同角色。其中一位仿生 (humanoid) 機器人更以講者的身分出現，分享其在長照上能做的事情。或許正如電影「機械公敵」一樣，機器人會逐漸擁有自己的意識，進而實施「保護人類計畫」。屬於我們這一代人的故事正在上演，留給我們準備的時間可能不多了。