

● 作者/Paul Scharre ● 譯者/周敦彥 ● 審者/黃依歆

人工智慧 軍備競賽危機

Killer Apps:
The Real Dangers of an
AI Arms Race

取材/2019年5-6月美國外交事務雙月刊
(*Foreign Affairs*, May-June/2019)

國際間對人工智慧軍備競賽的論調，會加速各國在安全上抄捷徑的行為。政策制定者一開始就應將安全性列入設計功能之內，並尋求與他國合作降低風險的機會，否則這場新科技競賽將不會有贏家。

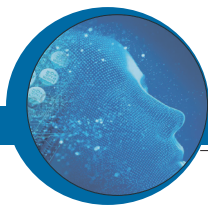
2017年，俄羅斯總統普丁表示，在人工智慧領域發展領先的國家「將主宰世界」。此一觀點是全球的共識，且已有超過十多個政府公布了人工智慧倡議。2017年中共設定目標，要在2030年成為人工智慧的全球領頭羊。今年稍早，白宮發表





數據蒐集、電腦處理能力和演算法設計的進步，讓研究人員以更靈活的人工智慧方法取得重大進展。

(Source: U.S. Department of Energy)



美國人工智慧倡議，而美國國防部也推出人工智慧戰略。

但是關於人工智慧軍備競賽新論述反映出人工智慧風險的錯誤觀點，並因此帶來全新的重大風險。對於每個國家而言，真正的危機並不是在人工智慧領域中落後於競爭對手，而是對軍備競賽的認知會促使各國急於運用不安全之人工智慧系統。在求勝的企圖中，每個國家所面臨之風險和對手是一樣大的。

人工智慧很可能帶來極大的益處，從健康照護到交通運輸所有領域皆然，但也存在著巨大的風險。這些風險並非來自科幻小說；沒有必要害怕機器人叛亂。真正的威脅將來自於人類。

目前，人工智慧系統功能強大但並不可靠。許多系統難以防範精密的攻擊，或是在其訓練範圍外的環境使用時仍會產生故障。各國政府希望這套系統能正常運作，但是相互競爭下帶來投機的壓力。即使其他國家並沒有在人工智慧領域上有重大突破，但是因為認知到大家都在積極發展，自己也會跟著這樣做。如果政府運用未通過測試的人工智慧武器系統，或是藉由有缺陷的人工智慧系統來發動網路攻擊，那麼可能會對所有相關人員都造成災難性後果。

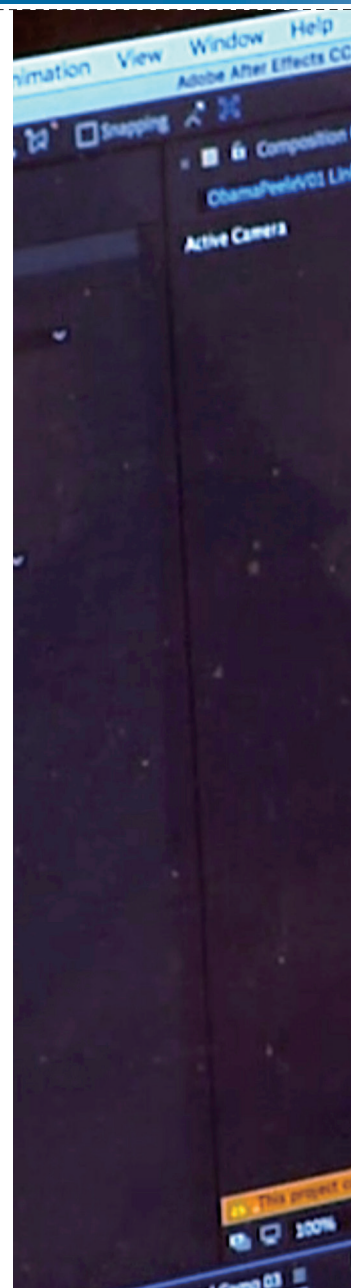
政策制定者應該從電腦網路的歷史汲取教訓，並從一開始就將安全性列入人工智慧設計的主要因素。他們應該減少關於人工智慧軍備競賽的言論，並尋求與其他國家合作的機會，以減少人工智慧的風險。競相降低人工智慧的安全門檻會是一場沒有贏家的競賽。

規則性的人工智慧系統

最簡易的人工智慧系統是依循人類預先設定一系列規則來執行任務。眾所周知，這些「專家系統」已存在數十年。這些規則無所不在，以至於一般人幾乎不會想到飛機自動駕駛儀或報稅軟體背後的技術就是人工智慧。但在過去幾年中，數據蒐集、電腦處理能力和演算法設計的進步，使得研究人員以更靈活的人工智慧方法取得重大進展：機器學習(machine learning)。

在機器學習中，程式設計人員不用編寫規則；而是機器藉由分析被給與的數據來自行學習。提供演算法數千個已標記的物體照片，它就能學習聯結圖像規則與物體名稱。目前人工智慧熱潮始於2012年，當時研究人員使用稱為「深度學習」

(deep learning)的機器學習技術取得突破，該科技依賴深度神經網路。神經網路大致是一種模仿生物神經元的人工智慧科技，細胞與細胞之間溝通是藉由電脈衝的發送與接收。人工神經網路起初就像是一張對任何事物一無所知的白紙。系統藉由調適神經元之間連接的強度來學習，加強正確答案的某些路徑並削弱錯誤答案之連接。深度神經網路——負責深度學習的神經網路類型——是



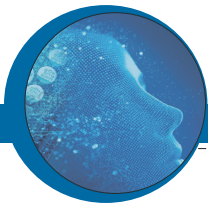


以人工智慧所產製的深偽影片，已經被用來進行報復性攻擊。(AP/達志)

在輸入和輸出層之間具有多層人工神經元的神經網路。更多層神經元在不同路徑強度上會有更多變化，並有助於人工智慧應對更多不同的情況。

系統學習的準確程度取決於機器學習演算法及開發人員使用數據的種類。許多方法使用已標記的數據(監督式學習)，但是機器也可以使用未

標記的數據(非監督式學習)或是直接從環境中學習(強化式學習)。機器還可利用合成、電腦產生的數據加以訓練。韋茂(Waymo)自動駕駛汽車公司旗下的自動車公路里程數已超過1,000萬哩，但該公司每天仍再以電腦模擬駕駛汽車行駛1,000萬哩，主要是為了在數十億的合成數據中測試其演



算法。

由於深度學習在2012年的突破，研究人員創造了各種人工智慧系統，在臉部辨識、物體識別、語音辨識及玩複雜遊戲各方面都能夠媲美或超越人類的最佳表現，其中包括華人的圍棋和美國線上電玩遊戲星海爭霸(StarCraft)。深度學習科技也開始超越老舊、規則式的人工智慧系統。2018年，深度學習演算法擊敗了國際棋賽電腦程式冠軍，之前僅花了四個小時在大型超級電腦上和自己對弈數百萬回合棋局，而沒有任何人類訓練數據或人工程式編碼規則來指導其行為。

研究人員目前正將人工智慧應用於一系列現實世界的問題，從診斷皮膚癌、駕駛汽車到提高能源效率。根據麥肯錫(McKinsey)顧問公司的估計，人們在美國獲得報酬從事的所有工作中，幾乎有一半可以透過現有科技實現自動化(雖然只有不到5%的工作可以完全被取代)。人工智慧工具也愈來愈普及。大型組織由於能夠累積大量數據並擁有強大電腦計算能力，最有可能取得重要技術突

破。但是許多人工智慧工具都可供所有人在線上取得。免費程式設計課程也教導人們如何製作自己的人工智慧系統，另外經過訓練的神經網路亦可免費下載。這些資源的易得性將可刺激創新，但將強大的人工智慧工具交到任何人手上也有可能助紂為虐。

專制的人工智慧

不當運用人工智慧的害處並非只是假設，而是已經發生。聊天機器人經常被用來操縱社群媒體，增加某些訊息的曝光度並壓制其他訊息。「深度偽造」(Deepfakes)係以人工智慧產生的假影片，已經被用於所謂的報復性色情攻擊，將人的臉部以數位化方式嫁接到色情演員的身上。

這些例子都只是開始而已。政治造勢活動將使用人工智慧數據分析，來針對特定人士進行量身訂製的政治宣傳。企業將使用相同的分析方式來設計操縱性廣告。數位竊賊使用人工智慧工具來發起更有效的網路釣魚攻擊。藉由複製人聲的一分鐘語音檔，聊天機器人將

能夠更逼真地在線上和透過電話扮演人類。任何不是親身的互動都會變得可疑。安全專家已證實，入侵自動駕駛汽車、使方向盤和剎車失效都是可能的。可想而知，一個人只要敲敲鍵盤就得以劫持整個車隊，造成交通堵塞或發動恐怖攻擊。

以人工智慧作為鎮壓的工具則更令人恐懼。威權政府可以使用深度偽造來詆毀異議者，透過臉部辨識實現全天候大規模監視，而預測分析則可識別潛在的麻煩製造者。中共已邁向數位威權主義的道路，並開始對新疆省的維吾爾族穆斯林進行大規模鎮壓活動。許多中共正在使用的工具技術性都不高，但他們也開始使用數據分析、臉部辨識系統及預測性警務(使用數據來預測犯罪活動)。透過龐大的監視攝影機網絡與演算法連結，可偵測異常的公眾行為，從違規停車到擅闖禁區都包含在內。中國大陸企業雲天勵飛科技自誇，中國大陸目前有將近八十座城市使用該公司的智慧攝影監視系統，並發現約6,000件和「社會治理」有關的事件。中共當局目前使



中國大陸目前有將近八十座城市使用雲天勵飛科技生產的智慧攝影監視系統，若一旦遭到當局濫用，將有侵犯人權之疑慮。(AP/達志)

用人工智慧的某些方式似乎無關緊要，諸如追蹤民眾在公共廁所使用多少衛生紙。但是他們未來可能的作法會更邪惡，例如監控用電模式以察覺可疑活動的跡象。

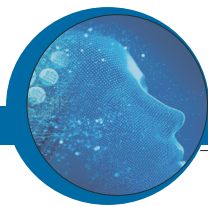
中共不僅在國內建立一個反科技烏托邦(tech-no-dystopian)監視國家，同時也已經開始輸出其科技。2018年，辛巴威與中國大陸雲從科技公司簽約，建立人臉國家數據庫，並在機場、火車站及公車站導入臉部辨識監視系統。該筆交易不僅只有涉及金錢。辛巴威已同意讓雲從科技將數百萬張臉部數據傳回中國大陸，幫助該公司改善針對深色皮膚人種的臉部辨識系統。中共還計畫在馬來西亞、蒙古及新加坡銷售監控科技。

中共也正在輸出其威權主義法律和政策。根據

自由之家(Freedom House)表示，中共與來自30多個國家的政府官員和媒體成員就監視與控制輿論的方法舉辦了訓練課程。有三個國家——坦尚尼亞、烏干達和越南——在與中共接觸後不久就通過了限制媒體及網路安全的相關法律。

人工智慧能做什麼

任何一個在人工智慧上領先的國家，都將會利用它來取得比競爭對手更多的經濟與軍事優勢。到了2030年，人工智慧預期將挹注全球經濟13兆至15兆美元。同時人工智慧還可以加快科學發展的速度。2019年，相較於生物學研究的關鍵工作——合成蛋白折疊(protein folding)的現有方法——人工神經網路的表現已明顯較佳。



人工智慧也將澈底改變戰爭。這很可能是改善士兵戰場狀況覺知、增進指揮官下達決心與傳達命令能力的最佳良方。人工智慧系統可以處理比人類更多的資訊，而且執行的速度更快，因此是即時評估混亂戰況的寶貴工具。在戰場上，機器運動的速度比人類更迅速，並且擁有更佳的精確度和協調性。最近人工智慧和人類在星海爭霸電玩遊戲的競賽中，AlphaStar人工智慧系統在快速處理大量資訊、協調並快速且準確地調度單元上展現了超越人類的能力。在現實世界中，這些優勢將使人工智慧系統比人類更有效地操控群集機器人。人類將在更高的戰略層次中保持優勢，而人工智慧則會主導實務應用。

中共已是人工智慧的全球霸主，華府擔心落後便急於發展人工智慧。中國大陸科技巨擘阿里巴巴、百度與騰訊，正好跟亞馬遜、谷歌及微軟並駕齊驅，名列全球領先的人工智慧公司。去年資本最雄厚的十家人工智慧新創企業中，就有五家來自中國大陸。早在十年前，中共誓言在2030年成為人工智慧全球領導者的目標看似天方夜譚；但如今真的已經成為可能。

同樣令美國政策制定者感到警覺的，是華府與矽谷在人工智慧的軍事用途方面存在巨大分歧。谷歌和微軟的員工反對其公司與美國國防部簽約，導致谷歌中止一項用人工智慧分析影片的計畫。而中共威權專制則不容許這種公開的反對意見。這種「軍民融合」的模式意味著中共科技創新更容易轉化為軍事用途。即使美國在人工智慧方面保持領先地位，也可能失去其軍事優勢。面對另一個國家贏得人工智慧競賽所構成的威脅，

合理反應是加倍本身對人工智慧的投資。問題在於人工智慧科技不僅威脅到這場競賽的輸家，也給那些贏家帶來風險。

新科技之弱點

今天的人工智慧科技功能強大但不可靠。以規則為運作基礎的系統無法應付程式設計人員沒有預料到的情況；而學習系統也受到所受訓練數據的限制；人工智慧的失敗已經造成悲劇。汽車的先進自動駕駛功能雖在某些情況下運作良好，卻在沒有警告的情況下駛往卡車、明顯路障及停置車輛的方向。在錯誤的情況中，人工智慧系統會在瞬間從絕頂聰明變成超級笨蛋。當敵人試圖操縱和入侵人工智慧系統時，這樣的風險甚至會更大。

學習系統即使未完全崩潰，有時也會以錯誤的方式習得達成目標。在2018年的一篇研究論文中，一個有52名人工智慧研究員的團隊，計算數十次人工智慧系統表現出令人驚訝的行為。一個學習在模擬環境中行走的演算法，發現它本身可以藉反覆摔倒而移動得最快。一個玩俄羅斯方塊的機器人學會在最後一塊磚落下前暫停遊戲，這樣就永遠不會輸。某個程式刪除了評量答案的檔案，使其獲得滿分。正如研究員們所稱，「為實現發展，實務上利用量化方法的漏洞通常比達成所欲結果更加容易。」出人意表似乎是學習系統的標準特徵。

機器學習系統的表現取決於訓練數據。如果數據無法充分體現系統的運作環境，系統就會在現實世界中失敗。例如，2018年麻省理工學院媒體

實驗室(Media Lab)研究人員證實，三個主要臉部辨識系統在分辨深色皮膚臉部的能力上遠比分辨淺色皮膚差得多。

機器學習系統在運作失敗時也常常令人沮喪難懂。對於以規則為運作基礎的系統而言，研究人員總能解釋機器的行為，即使無法隨時預測。然而，對於深度學習系統，研究人員往往無法理解機器為什麼會如此運作。谷歌人工智慧研究員拉希米(Ali Rahimi)認為，這就很像中世紀煉金術士一樣，他們發現了現代玻璃製造技術，但卻不了解背後化學或物理學原理，現代機器學習工程師可以取得豐碩的成果，但缺乏基礎科學來解釋它們。

人工智慧系統運作失敗也暴露其可利用的弱點。在某些情況下，攻擊者能夠破壞訓練數據。2016年，微軟創造一位名為泰伊(Tay)的聊天機器人，並給了它一個推特帳戶。其他用戶開始對其帳戶發布攻擊性推文，24小時內，泰伊開始模仿他們的種族主義和反猶太主義言詞。在此案例中，不良數據的來源是顯而易見的。但並非所

有毒害數據(data-poisoning)攻擊都如此明顯。有些雖以人類無法察覺的方式隱藏在訓練數據之中，仍具有操縱機器的能力。

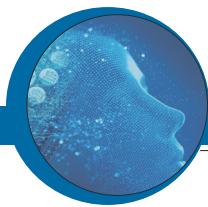
即使深度學習系統的創造者有保護其數據來源，系統仍然可能被所謂的「對抗樣本」(adversarial examples)欺騙，攻擊者向系統提供精心設計的資料，蓄意誤導機器出錯。鑑別衛星影像的神經網路可能會被欺騙，將巧妙修改的醫院影像誤認為是軍用機場，反之亦然。影像的變化能細微到人眼無法辨識，也能成功騙過人工智慧系統。對抗樣本甚至可以置於實際物品中。在一個案例中，研究人員創造了一個塑料烏龜，在龜殼中嵌入精細的漩渦，使物體識別系統誤認為是步槍。在另一個案例中，研究人員將一些白色和黑色的小方塊放在停車號誌上，導致神經網路誤認為是每小時45哩的限速標誌。更糟糕的是，攻擊者可以開發出這些欺騙性的圖像和物品，而無須侵入被攻擊系統的訓練數據或基礎演算法，因此研究人員一直在努力尋找有效防禦威

脅的方法。與網路安全漏洞不同的是，尚無可在被發現時進行修補，進而完全預防這些攻擊的演算法。

政府已具有測試軍事、網路和監視工具的豐富經驗，但卻無任何測試方法可保證這套複雜的系統在現實世界中不會故障。F-22戰鬥機首次飛越國際換日線時發生電腦當機，而飛機幾乎滯留在太平洋上空。

測試人工智慧系統通常比測試傳統軍事硬體要花費更多的時間和金錢。複雜性使其能力更為強大，但發生意外故障的機率也愈高。想像一下，政府開發一款人工智慧系統，可在不被發現的狀況下侵入對手的電腦網路。第一個運用這種系統政府將獲得巨大的優勢。由於擔心對手正在開發類似的工具，政府可能會感到不得不縮短測試時間並儘早部署這套系統。這種動態狀況早已發生在其他產業，例如自動駕駛汽車。然而，國家安全人工智慧工具所引起的意外後果應會更為嚴重。

人工智慧並不是政府所依賴之強大但危險的第一項科技。



電腦就是個例子，即便存在巨大的弱點，其在所有事情中都扮演重要的角色，從股票交易到導彈發射無所不包。2018年，美國政府責任署(U.S. Government Accountability Office)調查人員發現，美國的武器系統充斥著可能會被「相對簡單的工具和技術」利用的網路安全漏洞。更嚴重的是，國防部的程式負責人不知道這些問題，聲稱這些測試不切實際並駁回了該署的調查結果。電腦安全漏洞不僅限於政府系統。一家又一家的公司遭遇重大數據洩露危害。數位安全已長期受到忽略。一個遍布未受保護人工智慧系統的世界不僅僅是可能而已，而是註定會發生。

安全第一

緊急的威脅需要緊急的反應。政策制定者因應人工智慧危機最重要方式之一，就是增加人工智慧安全研究的經費。民間公司花費數十億美元尋求人工智慧商業應用，但美國政府可在資助基礎人工智慧研究方面扮演要角，就像早期一樣。由美國防先進研究計畫局

(Defense Advanced Research Projects Agency, DARPA)推動的次世代人工智慧(AI Next)倡議，打算在未來五年內投入20億美元，旨在解決狹義人工智慧系統的許多侷限性。為了擴大成果，白宮應增加人工智慧安全研究的經費，作為新美國人工智慧倡議(American AI Initiative)的一部分，並且應該向國會申請更多的研發和安全研究資金。

在將人工智慧應用於國家安全方面時，政府機構必須重新考慮其測試新系統的傳統方法。驗證系統是否符合其設計規範是不夠的。測試人員還需要確保當敵人試圖擊敗它時，依舊能夠在現實世界中持續正常運作。在某些情況下，他們可以使用電腦模擬來挑出錯誤，正如目前自動駕駛汽車製造商的作法。最重要的是，美國國防部與國土安全部及情報體系應該創建敵軍團隊(red teams)——扮演攻擊者來測試的防禦能力——找出人工智慧系統的弱點，以便開發人員可在系統上線之前予以修正。

政府官員也應淡化關於人工

智慧軍備競賽的言論，因為這樣的談話很容易變成自我實現。在2018年的一次會議中，五角大廈主管研究和工程的官員葛里芬(Michael Griffin)表示，「可能會有一場人工智慧軍備競賽，但我們尚未參與其中。」軍方肯定會運用人工智慧，但是葛里芬的聲明沒有提及——或甚至意識到——任何隨之而來的風險。談論軍備競賽促使敵人在安全方面抄捷徑。政府官員不僅要強調人工智慧的價值，還要強調確保可靠性及安全性的重要。

最後，美國應該尋求和其他國家合作的方式，甚至是與敵對國家聯手確保人工智慧安全。新科技的國際合作方案在過去有各種結果，但各國間共同努力成功避免互傷亦時有所聞。在冷戰期間，美國和蘇聯共同合作限制了某些極不穩定的核彈頭發射系統類型。美國還鼓勵其他國家採取安全措施，防止未經授權使用核武。今日，美國應與盟國和敵人合作，增加國際發展人工智慧安全的資金，同時更應開始與中共和俄羅斯討論，是否某些人工智慧

應用會引起不可接受的衝突升高或失控風險，以及各國可以共同採取哪些措施來強化安全性。美國在人工智慧競賽中面臨的最大危機不是失敗，而是創造了一個沒有贏家的世界。

十九世紀時，工業化帶來經濟大幅成長，但也讓軍隊擁有戰車、機槍和芥氣。核武的發明帶來更深沉的風險，政策制定者迄今仍努力解決此一難題。電腦澈底革新人們工作、學習和溝通的方式，但也導致過去孤立的系統易遭受網路攻擊。

人工智慧相較於以上科技變化猶有過之。但大多數將是正面的影響。它會促進經濟增長、有助診斷和治療疾病、減少汽車事故，並以大小小

等數千種方式改善人們的日常生活。然而，就像任何新科技一樣，人工智慧也有黑暗的一面。現在就勇於面對風險，是確保人類實現人工智慧前景而非釀成禍害的唯一途徑。

作者簡介

Paul Scharre係新美國安全中心(Center for a New American Security)資深研究員和科技與國家安全計畫主任。他著有《無人軍隊：自主武器與未來戰爭》(Army of None: Autonomous Weapons and the Future of War)一書。

Copyright © 2019, Council on Foreign Relations, publisher of *Foreign Affairs*, distributed by Tribune Content Agency, LLC.



電腦澈底革新人們工作、學習和溝通的方式，但也導致過去孤立的系統易遭受網路攻擊。(Source: AP/達志)