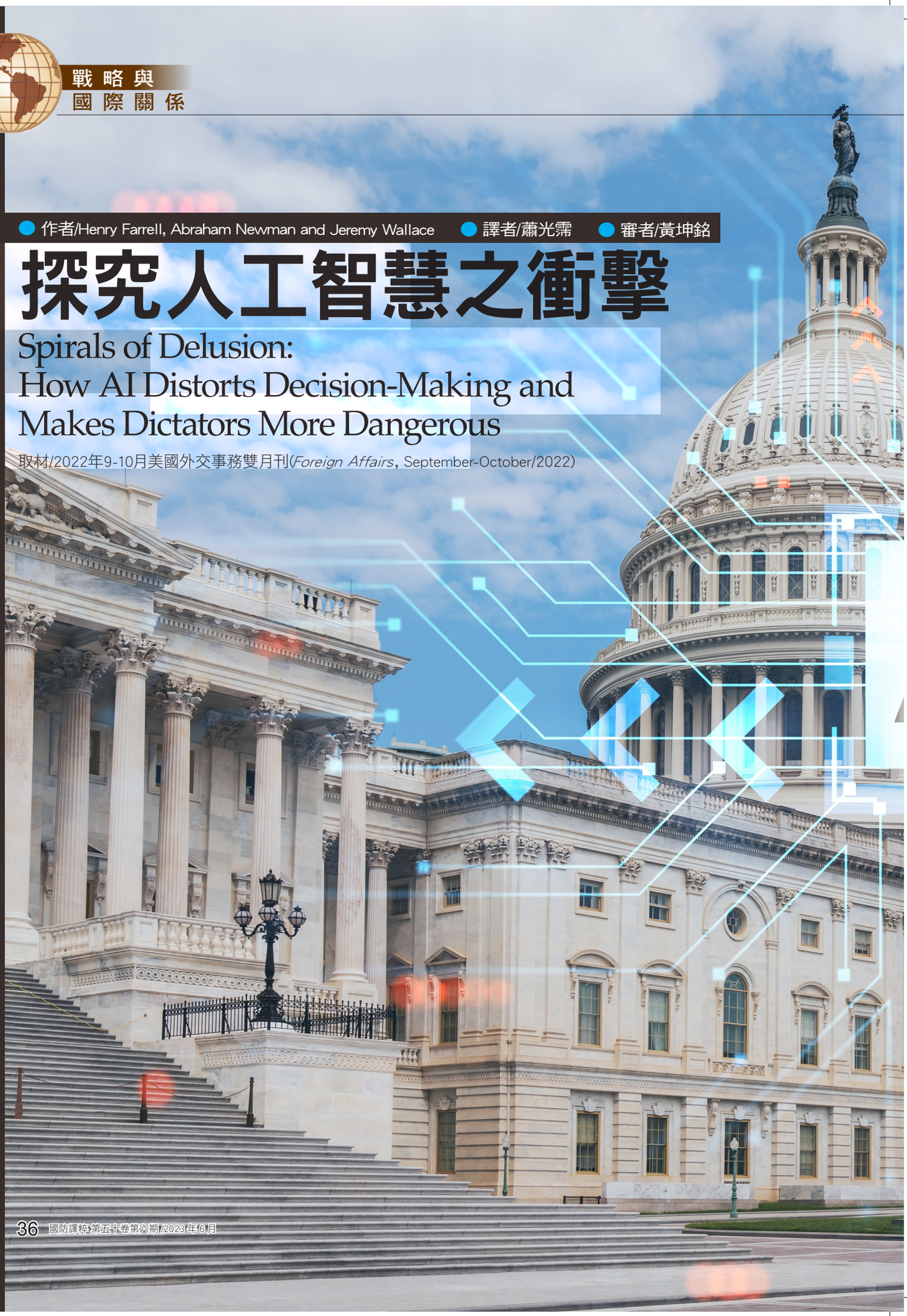
作者/Henry Farrell, Abraham Newman and Jeremy Wallace　譯者/蕭光霈　審者/黃坤銘

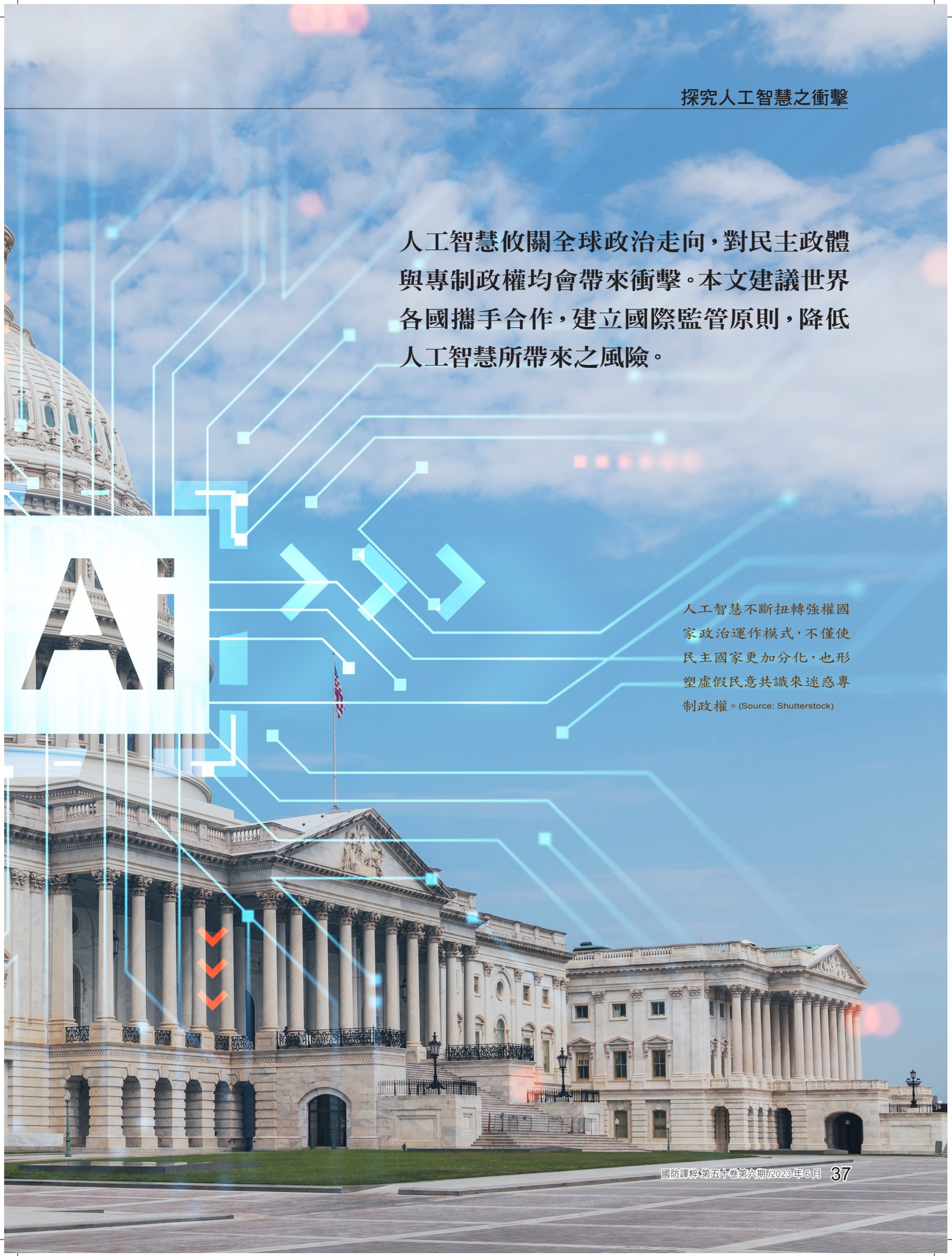
探究人工智慧之衝擊

Spirals of Delusion: How AI Distorts Decision-Making and Makes Dictators More Dangerous

取材/2022年9-10月美國外交事務雙月刊(*Foreign Affairs*, September-October/2022)



人工智慧攸關全球政治走向，對民主政體與專制政權均會帶來衝擊。本文建議世界各國攜手合作，建立國際監管原則，降低人工智慧所帶來之風險。



人工智慧不斷扭轉強權國家政治運作模式，不僅使民主國家更加分化，也形塑虛假民意共識來迷惑專制政權。(Source: Shutterstock)

政界人士談到人工智慧，總是聚焦中共與美國競逐科技霸權。若關鍵資源是數據，中共擁有十多億人口，且防範國家監控做法鬆散，似乎勝券在握。資深電腦科學家李開復曾指稱，數據是新石油，中共等同於全新「石油輸出國家組織」。若科技領先能夠形成優勢，美國擁有世界聞名大學體系與傑出勞動人力，仍有機會占上風。權威人士認為，兩國之間，孰能取得人工智慧優勢，經濟與軍事優勢必會水到渠成。

但從競逐霸權的角度來思考人工智慧時，將會忽略人工智慧如何根本扭轉全球政治。與其說人工智慧扭轉強權對峙態勢，倒不如說對敵對強權國家自身產生更大衝擊。美國是民主政體，中共是專制政權，而機器學習正不斷改變兩個政治體制。美國這種民主政體所面臨之挑戰顯而易見。機器學習讓國家意識形態更加兩極化，如網路言論讓國內政治更加分化。機器學習未來當然更會散播假訊息，創造大量危言聳聽之虛假言論。機器學習對於專制政權之挑戰則更為微妙，但卻更深具破壞力。就如同機器學習凸顯與加深民主政體內部分化，其亦能形塑某種共識表象以迷惑專制政權，隱瞞潛藏社會分裂，直到最終回天乏術。

包括政治科學家司馬賀(Herbert Simon)在內的人工智慧早期先驅，皆明瞭人工智慧技術不僅是單純工程應用，其與市場、官僚體系及政治制度有諸多雷同。另一位人工智慧先驅維納(Norbert Wiener)係將人工智慧形容為「模控化」(Cybernetic)系統──能對變化做出反應，並予以反饋。但司馬賀與維納皆未預期機器學習如何主導人工智慧發展，惟機器學習演進符合前述兩位學者預期。臉書與谷歌使用機器學習做為自我校正系統之分析引擎，該系統會依據過往預測結果，持續優化數據分析方式。就是此種統計分析與環境反饋循環，讓機器學習變得勢不可當。

然外界多半不瞭解，民主與專制亦屬「模控化」體系。在兩種治理方式下，政府施行政策，然後設法探究施政良窳。在民主政體中，選票與民意即為施政良窳之反饋。在歷史上，專制體系較難取得正向施政反饋。資訊時代來臨前，專制體系依賴國內情蒐、人民請願及暗中調查輿情，瞭解民心所向。

今日，機器學習正在裂解民主反饋(民意與選票)傳統形式，因為新科技能將偏見暗藏於數據中，悄悄轉化為錯誤主張，因而助長假訊息，加深成見。對於在暗中摸索的獨裁者而言，機器學習似乎恰可實現所願。這種技術能讓統治者不必耗費心力調查輿情，冒著政治風險舉行公開辯論與選舉，就能瞭解臣民對其施政滿意程度。因此，許多觀察家憂心，人工智慧逐漸進步，只會強化獨裁者鐵腕作為，讓他們更能夠控制社會。

然而事實更加複雜。偏見在民主政體中顯而易見。由於偏見清楚可辨，人民可透過其他反饋以緩和其所帶來之衝擊。例如，特定人種族群認為招聘辦法有歧視之虞，就會進行陳抗、試圖改正並予以補救。專制政權內部偏見問題可能與民主政體不相上下，也許更多。然而，該政權內部偏見可能大多難以察覺，高層決策者更不易發現。因此，即便領導者心知肚明，偏見也難以導正。

與傳統思維相反，人工智慧能夠強化專制政權意識形態與自我陶醉，進而與現實世界脫軌，走

火入魔。民主政體應能發覺，談到人工智慧，贏得科技霸權並非二十一世紀關鍵挑戰。民主政體反倒要與深陷人工智慧虛假螺旋中的專制政權一較高下。

惡性循環

探討人工智慧多半都會涉及機器學習，機器學習就是透過統計演算分析數據間之關聯。這些演算法可進行推測：相片中有狗嗎？這套棋法可在十步內得勝嗎？句子寫到一半，下個字為何？另一個所謂「目標函數」(Objective Function)，即獲取結果之數學演算法；若演算法推測正確，「目標函數」就能再精進此演算法。此即商業上人工智慧運作過程。舉例而言，YouTube期望吸引用戶觀賞更多影片，藉此提高廣告收視率。「目標函數」意在提高用戶參與程度。演算法會試著提供吸睛內容，讓用戶目不轉睛。演算法依據用戶喜好，調整影片類型，提高用戶黏著度。

機器學習能夠在些許或毫無人類介入情況下，自動完成前述反饋流程，此舉完全翻轉過往電子商務運作模式。未來，機器學習可能全自動駕駛車輛，縱使目前發展碰到工程師預料之外的瓶頸。發展自主武器仍是艱澀課題。當演算法遭逢亂流，面臨意外事件，通常就無法理出頭緒。雖然人類可輕易瞭解，但導致機器學習歸類錯誤的資訊──或稱為「對抗樣本」(Adversarial Example)，會讓系統運作遭受重大阻礙。舉例而言，在停止號誌上黏貼黑白相間貼紙，自駕車視覺系統就無法辨識。此類弱點顯現人工智慧戰時運用限制。

探究機器學習的複雜性，有助於瞭解科技霸權相關論辯。此舉亦說明，為何部分思想家(如前述電腦科學家李開復)認為數據如此重要。數據愈多，就愈能快速改善演算法效能，不斷微調以獲取決定性優勢，但機器學習有其限制。例如，儘管科技公司挹注大量資金開發演算法，然演算法遠不及傳聞所言，能在民眾對兩種幾近相同產品中猶豫不決時，影響買家選購特定種類。百分之百操弄消費習慣殊為不易，若要改變民眾根深蒂固想法與信念，恐更是難上加難。

通用人工智慧(General AI)與人類相仿，可從某個情境中汲取經驗，並運用在其他情境，但亦面臨前段所提及之類似限制。網飛(Netflix)統計用戶喜愛與偏好，但分析同類型顧客對眼前商品猶豫不決所採用的分析模型，也與亞馬遜(Amazon)全然不同。在人工智慧某一領域取得優勢──例如能提供讓年輕人目不轉睛的短片(如抖音應用程式大受歡迎)，無法輕易轉換為另一領域優勢，就拿打造自主武器系統做為比方。演算法成功與否，通常著重人類工程師從不同應用程式運作中，汲取經驗並予以轉化，而非由科技本身自主轉換。前述問題迄今仍懸而未決。

偏見亦會潛藏在程式碼中。亞馬遜曾嘗試運用機器學習進行招募，從招募員審閱過的人資中擷取數據，續由演算法執行運算。結果卻是重現人類決策固有偏見，對女性產生歧視，而這種問題會惡性循環。社會學家班傑明(Ruha Benjamin)指出，倘若政策制定者運用機器學習來決定警力部署地點，機器學習應會建議增派警力至逮

捕率高的社區，最終導致派遣更多警員至警方歧視種族群聚區域。此舉則會導致更多逮捕案件，讓機器學習演算法陷入惡性循環。

編寫程式上有句俗諺「垃圾進，垃圾出」，但在輸入影響輸出，反之亦然的環境下，這句俗諺則擁有不同意涵。機器學習演算法在無外力適切校正的情況下，可能會對垃圾產出產生偏好，形成一個惡性決策循環。政策制定者通常將機器學習視為是充滿智慧與沉著冷靜的專家，而非原本要用來解決問題，最後卻雪上加霜、錯誤百出的工具。

一呼百應

政治體系亦屬反饋體系。在民主政體中，民眾透過本應自由公平的選舉，評估領導者表現。政黨做出許多承諾，以贏得政權、尋求連任。合法反對勢力會揪出政府紕漏，自由媒體報導各項爭議與貪贓枉法之舉。在這樣周而復始循環中，在位者經常要面對選民，以瞭解能否贏得民眾信任。

但在民主社會中，民眾反饋成效不盡理想。人民也許對政治瞭解不深，人民可能針對政府能力所不及之事項予以譴責。政治人物與其幕僚恐不瞭解民之所欲，而反對黨常會編造謊言、誇大事實。選舉所費不貲，而實際決策通常是在檯面下進行。媒體業也存在偏見，而且積極娛樂顧客，而非負起開導教化之責。

不變的是，有反饋才能促進瞭解。政治人物才能瞭解民之所欲，民眾才能抱持正確期待。民眾可以公開抨擊政府施政違失，而免於牢獄之災。當新問題浮出檯面，會有新團體成立，將問題公諸於世，並試圖說服他人解決問題。以上所述可讓政策制定者與政府就算身處龐雜紛亂且變化萬千的世界中，也不會與民意脫節。

反饋在專制政權中的運作方式極為不同。領導者不是透過自由公平選舉產生，而是透過殘酷的繼承爭鬥、晦暗不明的內部推舉制度產生。即便反對政府在形式上合法，但不受鼓勵，有時甚至會遭受殘暴壓制。若媒體批評政府，就會面臨法律制裁與暴力相向之風險。就算有辦理選舉，其制度亦有利於在位者。反對領導者的民眾不僅難以集結，若宣揚訴求，亦可能遭受入獄與死刑等嚴刑重罰。有鑑於此，專制政府通常對世界運作方式或民之所欲存在錯誤認知。

因此，專制政權必須權衡短期政治穩定與有效政策制定兩者間之輕重。若是在意短期政治穩定，專制領導者就會阻擋外界表達政治意見；反之，若訴求有效政治決策，專制領導者就須瞭解外界與內部社會現況。由於嚴格控管資訊流通，專制領導者無法像民主政體領導者一樣，依賴民眾、媒體及反對聲浪提供真實反饋。結果是專制領導者要冒著政策失敗，冒著葬送長期政權合法地位與統治能力之風險。例如，俄羅斯總統蒲亭入侵烏克蘭的災難性決定，似乎就是基於對烏克蘭士氣與俄羅斯自身軍力之錯誤評估。

即使在機器學習問世前，專制領導者就運用原始粗淺且未盡周延的量化做法來替代民眾反饋。以中共為例，中共數十年來試著將分散市場經濟結合集權政治監管，掌握如國內生產毛額等少數重要統計數據。若地方官員所轄區域經濟有顯著成長，就能加官進祿。但北京當局這種侷限量化觀點，讓地方官員無意解決如貪汙、債務與汙染等

惡化中之問題。可想而知，地方官員通常會玩弄統計數字，或採取能在短期拉升國內生產毛額之政策，而長期問題則留給接任者傷腦筋。

2019年底，中共對湖北省新冠肺炎案例之初期反應，讓世界各國得以一窺其發展。2003年，中共於嚴重急性呼吸道症候群(Severe Acute Respiratory Syndrome, SARS)危機後建立一套網路疾病通報系統，但湖北省首都武漢市官員卻未使用該系統，反而處分第一時間通報出現「類嚴重急性呼吸道症候群」傳染病的醫師。武漢市政府努力阻止疫情爆發的訊息洩漏至北京當局，在當地重要政治會議結束前，不斷重申「無新增病例」。而該名醫師李文亮則於2020年2月7日不敵病情身亡，當時引起民眾強烈不滿。

於是，北京當局接手傳染病應變作為，採取「清零」作法，使用高壓手段來壓低確診人數。這項政策短期內運作良好，但在Omicron變異株強大傳染力席捲下，「清零」政策可能只是兩敗俱傷，採取大規模封控作法只會致使人民挨餓、經濟受創。但此舉仍算成功達到一項重要指標：維持低確診人數。

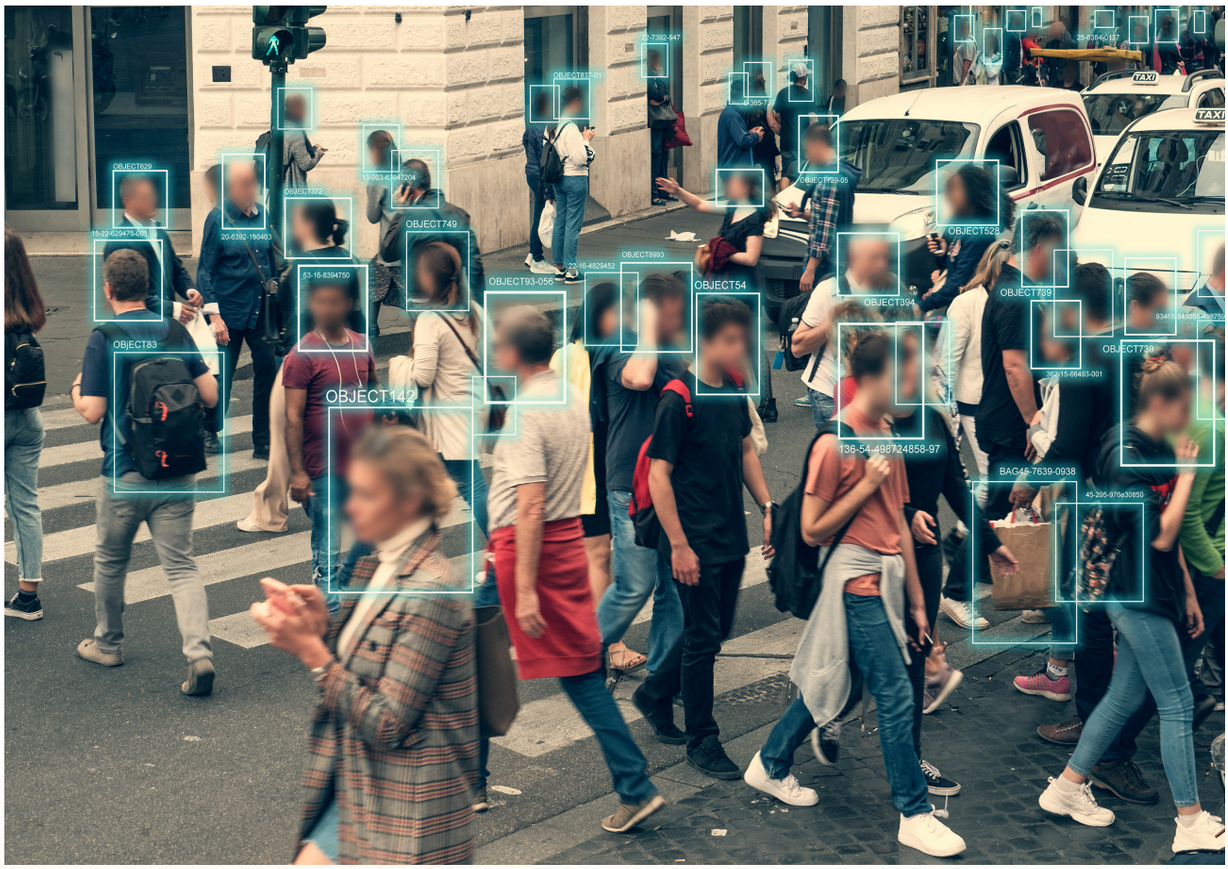
在不冒政治風險與避免選舉或自由媒體構成不便之情況下，數據似可做為詮釋全世界與其問題的客觀做法。然而，拋開政治的決策作為根本就是空談。對於關注美國政治的民眾而言，不難發現民主政體紛亂與反饋過程脫序的風險。專制政權亦面臨類似問題，雖然這些問題較無法立即察覺。官員在數字上造假或人民不願將憤怒轉化為大規模抗議，皆會構成嚴重後果：短期內可能會做出錯誤決斷，日積月累之下恐導致政權垮臺。

請君入甕？

當下最迫切的問題，並非美「中」誰會取得人工智慧霸權，而是在於人工智慧如何轉變民主政體與專制政權賴以治理國家之反饋循環。許多觀察家指出，當機器學習日漸普及，將無可避免地傷害民主政體、助長專制政權。從他們的觀點來說，例如能夠優化人類接觸交流的社群媒體演算法，恐有損民意反饋品質，進而傷害民主政體。當民眾不斷瀏覽影像，Youtube演算法就會一直提供聳動與吸睛內容，讓人目不轉睛。而這些內容通常涉及陰謀論或極端政治觀點，吸引民眾進入顛倒是非的黑暗國度。

相形之下，機器學習讓專制政權更容易操控民眾。歷史學家哈拉瑞(Yuval Harari)與其他學者主張，人工智慧「對暴政有利」。依據前述學者主張，人工智慧將數據與權力集中，讓領導者能精打細算，提供觸動人民「心弦」的訊息，操控百姓思維。藉此無窮無盡、周而復始的反饋與回應過程，潛移默化、成效卓著地控制社會。因此，社群媒體協助專制政府瞭解人民，捕捉民心。

但前述主張係建構於毫無事實之基礎。雖然從臉書公司外流的訊息指出，演算法確實可誘導民眾瀏覽激進內容，近期研究卻指出演算法本身並未改變民眾心之所向。搜尋YouTube偏激影片者的民眾，可能基於本身所需而接受演算法推薦內容；但對危險內容不感興趣的民眾，就未必會採納演算法的建議。若在民主社會中之反饋變得愈來愈脫序，機器學習不應成為千夫所指，機器學習能夠做到的僅



專制政權領導者嚴格管制資訊流程，透過人臉辨識掌握民眾行為，卻無法如實掌握民意向背。(Source: Shutterstock)

不過是「助長」此狀況。

目前尚無足夠證據顯示，機器學習能達到掏空民主、強化專制之普遍性心智控制。若演算法無法有效改變民眾消費意向，遑論動搖民眾堅信之價值觀(如政治觀點)。2016年，英國政治顧問公司劍橋分析(Cambridge Analytica)運用某種神奇技巧為川普(Donald Trump)於總統大選舞弊之傳言已甚囂塵上。該公司提供川普打選戰之秘密方程式，似乎包括通用性有限心理計量標定法(Psychometric Targeting Technique)，亦即依據個性將選民分類。

誠然，對於中共這種權力集中於少數與世隔絕決策者的國家，全自動化、數據導向之專制主義恐怕是個陷阱。民主政體具有矯正機制，亦即當政府脫軌時有人民的反饋機制予以制衡。專制政府加倍運用機器學習，卻無前述矯正機制。雖然如影隨形的國家監控短期內可能有所成效，然其危險之處，在於專制政權會遭自我強化與機器學習推波助瀾下形成之偏見所反噬。當一個國家廣泛使用機器學習，其領導者之意識形態

會形塑機器學習運用方式、機器學習優化資料之最終目標，以及機器學習對於結果之詮釋。依此過程產生的數據，恐怕會反映領導者個人偏見，讓自身成為眾矢之的。

科技界人士賽格洛斯基(Maciej Ceglowski)認為，機器學習似在「為偏見洗錢」，係為一「簡潔之數學裝置，能讓現狀在邏輯上陷入死胡同」。舉例而言，當國家開始運用機器學習尋找社群媒體上的牢騷埋怨並將其根除，則後果為何？就算政策錯誤危及政權，領導者也不易發現，難以解決實質問題。2013年更有研究推測，中共移除網路上批評埋怨言論的速度比起外界預期更慢，正因為強烈的批評埋怨，係為領導者提供真知灼見。但目前北京當局愈來愈強調社會和諧，並意圖保護高層官員，過往放任網路言論之做法勢必難以為繼。

中共國家主席習近平瞭解這些問題，至少對某些政策層面窒礙有所掌握。習近平長期主張脫貧──消弭偏鄉貧苦──就是運用智慧科技、大數據及人工智慧所獲得之指標性勝利。但習近平瞭解此舉也造成反效果，其中包括官員逼迫百姓離開偏遠家鄉，搬到都市公寓中人擠人，藉此玩弄貧窮統計數字。此刻，移居者再度跌回赤貧，習近平憂心，貧窮階層之「制式量化目標」恐怕不適合當作未來脫貧施政成效指標。數據可能的確是新石油，但亦恐拖累(而非強化)政府治理能力。

此項問題對於中共所謂之社會信用體系(Social Credit System)有其意義，社會信用體系係用以追蹤支持社會行為的制度，被西方評論家形容為「由人工智慧推動、違反人權的監控制度」且功能健全。如資訊政治專家阿邁德(Shazeda Ahmed)與郝珂靈指稱，事實上該體系相當雜亂無章。中共社會信用體系其實更像是美國受到《公平信用報告法》(*Fair Credit Reporting Act*)監管的信用體系，而不像是完美的歐威爾式反烏托邦(Orwellian Dystopia)。

過度運用機器學習，恐讓專制政權更容易做出不良決策。若機器學習經過訓練後，能從逮捕紀錄中找出潛在異議分子，機器學習就會強化自我偏見，如同民主政權所見一斑，反映並確立執政當局對於部分社會團體之偏見，不斷產生懷疑與激烈反彈。民主政體中，民眾還有可能公開對政策表達反對意見。然在專制政權中，民眾更難抵抗；沒有抵抗，體制內官員與演算法皆抱持相同偏見，內部人士就會對問題視而不見。此舉不會形成良好政策，而只會讓政權更加千瘡百孔、社會失能、民怨四起，終究造成社會動盪與不安。

人工智慧武器化

從國際政治角度來看，人工智慧並不只是霸權競賽。將人工智慧技術視為是透過數據所驅動的一種經濟與軍事武器之粗淺看法，其中暗藏許多實際舉動。事實上，人工智慧在政治上的最大影響，顯現在民主與專制政權所依賴的反饋機制。某些證據顯示，人工智慧正在擾亂民主政體的反饋機制，就算許多人認為人工智慧尚未如此舉足輕重。相形之下，專制政府愈依賴機器學習，愈會讓本身陷入科技強化偏見所建構的虛幻世界中。1998年，政治學家

史考特(James Scott)經典名著《國家的視角》(*Seeing Like a State*),解釋二十世紀國家如何對其所作所為視若無睹,部分原因是這些國家只透過政治體制與數據這兩個面向來透析這個世界。社會學家弗凱德(Marion Fourcade)與其他人士辯稱,機器學習恐重演歷史,但影響規模恐將更大。

此項問題對於美國這種民主政體,構成不同面向的國際挑戰。舉例而言,俄羅斯投入經費,展開對國內民眾散播製造紛亂與不安的假訊息活動,同時亦以相同伎倆對付民主政體。雖然致力維護言論自由人士長期主張,對抗惡劣言論的辦法,就是發表更多實況報導,而蒲亭卻認為要發表更多惡劣言論來抹黑前述實況報導。然後,俄羅斯利用民主政體開放反饋體系,散布錯誤訊息混淆視聽。

其中一個迫切問題,就是俄羅斯這種專制政權會將大型語言模式武器化──此係新型態人工智慧,可依據言語提示產生文字或影像,再產出大量假訊息。電腦科學家格布魯(Timnit Gebru)與同僚提出警告,開放人工智慧(Open AI)公司GPT-3系統已能產出相當流利、真假難辨之文字,與人類寫作內容相仿。開放大型語言模型Bloom才剛公開大眾使用,雖然使用授權要求使用者避免濫用,但未來監管仍相當困難。

前述發展將對民主政體構成嚴重問題。當前線上政策評論系統幾乎皆會遭殃,因為這些系統對於評論者身分幾乎不予驗證。大型電信公司之合約商已採取行動反對《網路中立法》,使用遭竊電子郵件地址寄出假評論,塞爆美國聯邦通訊委員會(Federal Communications Commission)網站。然而,當數以萬計內容幾近相同的評論貼文湧現,這種伎倆很容易被看穿。現在或在不久的將來,以某種大型語言模型,用中間選民口吻,寫出二萬分不同評論譴責《網路中立法》,根本就是易如反掌。

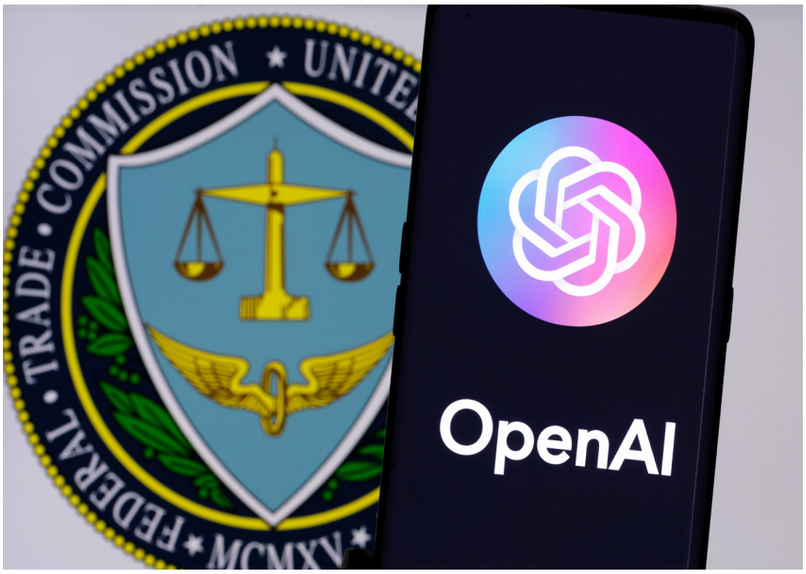
人工智慧推波助瀾的假訊息,亦能為專制政權到處扣帽子。專制政府公開散播假訊息,很容易就能分化反對勢力,但也更難瞭解民眾真實想法,讓政策制定過程更加複雜。這樣一來,專制領導者更容易閉門造車,認為人民能容忍,甚或接納極不受歡迎的政策。

共同威脅

中共這種專制政權深陷本身不健全的訊息反饋循環,那若與這類國家在世界上共處的結果為何?若上述反饋機制不再提供「模式化」指導,如實反映統治者內心恐懼與想法時,會產生什麼後果?民主競爭者若以自我為中心,恐怕就會讓專制者自行了斷,放任專制政府日漸衰敗,坐享其成。

然而,如此回應將會導致人道危機。中共政府目前諸多偏見(如維吾爾政策)都造成不少危害,未來恐怕會更加惡化。北京當局先前對實際狀況視若無睹之後果,導致1959至1961年大饑荒,造成三千餘萬人喪生,後續也因意識濃厚之政策加劇災情,而省級官員也不願如實回報確切確診人數,試圖隱匿疫情。即便是偏執的憤世嫉俗人士也能瞭解中共與其他國家因人工智慧引起之外交政策災難。舉例而言,人工智慧輕而易舉就可加深民族主義,鞏固鷹派勢力。

也許更諷刺的是,西方政策



美國及其盟國應瞭解人工智慧潛在風險，共同擬定相應對策，避免美「中」間擦槍走火。(Source: Shutterstock)

制定者或許也想利用專制政權資訊體系上封閉式回饋機制。目前，美國致力在專制社會中提倡網路自由。美國現在反而可能投其所好，設法藉由強化專制政權現有偏見循環，惡化專制政權的資訊問題。美國可藉破壞行政數據，或提供專制政權官媒假訊息，以達成這項目的。不幸的是，民主與專制體系並非涇渭分明。劣質數據與瘋狂想法，不僅會從專制社會滲入民主社會，專制政權之拙劣決策亦恐對民主政體構成嚴重後果。當政府面對人工智慧時，必須切記，我們生活在相互依存的世界，專制政府的問題亦會發散到民主政體中。

然而，較為明智的做法是透過國際治理制定共同協議，消弭人工智慧的弱點。目前，中共不同部門對於監管人工智慧抱持不同意見。舉例而言，中共國家互聯網信息辦公室、信息通信研究院與科學技術部皆針對監管人工智慧擬定相關原則。某些單位支持由上而下進行監管，藉此限制民間企業，並維持政府監管彈性。其他單位雖未明確表述，但亦瞭解人工智慧對政府構成威脅。研擬共同國際監管原則，應有助於述明人工智相關政治風險。

此種糾合群力之舉，在美「中」競爭日益升溫的情況下，似乎有違常理。但透過密切協調而擬定之政策，對美國與其盟國有所助益。若美國捲入人工智慧爭霸賽，就會讓美「中」競爭愈演愈烈。不然，就是讓專制政權反饋問題更加惡化。但兩者皆會構成災難或引發戰爭。然而，更安全的做法是讓各國政府瞭解人工智慧之共同風險，並通力合作、攜手降低風險。

作者簡介

Henry Farrell 為美國約翰霍普金斯 (John Hopkins) 大學斯塔夫羅斯・尼亞爾霍斯基金會 (Stavros Niarchos Foundation) 國際事務教授。

Abraham Newman 係美國喬治城大學艾德蒙・沃爾什 (Edmund A. Walsh) 外交學院及政府學系教授。

Jeremy Wallace 為美國康乃爾大學助理教授，著有《尋找真相、隱匿事實：中共資訊、意識形態與專制主義》(*Seeking Truth and Hiding Facts: Information, Ideology, and Authoritarianism in China*，暫譯)。