《科際整合月刊》，2026 年 11 月，11(11)，47-93。
Interdisciplinary Research Monthly. November 2026

人工智慧何以成為國際政治的核心議題？

楊惟任*

摘 要

近年來，人工智慧（Artificial Intelligence, AI）逐漸超出產業與科技政策的範疇，成為元首外交、聯合國安全理事會（United Nations Security Council, UNSC）及多邊治理的重要議題。本文結合安全化理論、權力政治與全球治理研究，以「技術能力－競爭/合作轉化機制－焦點事件－核心議題化」說明其發展歷程，並提出議程層級、安全框架、資源戰略化、制度化密度、行為者擴散及跨議題連結六項判定指標。研究材料包括 2017 年至 2026 年 9 月的官方文件、國際組織紀錄、企業公開資料與新聞報導，採用文獻分析及過程追蹤。研究發現，AI 的通用目的性與代理自主能力，讓它既是國家競爭的權力資源，也是可能引發跨境風險的技術。國家透過戰略規劃、出口管制與安全論述提高 AI 的外交重要性；跨境風險則促使國家、國際組織及科技企業尋求共同規範。不同參與者分別以國家安全與人類整體安全為訴求，形成兩條安全化途徑。2026 年代理式 AI 越權事件接連曝光，與高階外交討論在時間和議題內容上有所呼應，符合焦點事件可能促成政策關注的預期。按照本文設定的六項指標，AI 於 2026 年符合「核心議題」的判定條件；但事件調查尚未完成，門檻設定也有待跨議題檢驗。本文據此討論事故通報、獨立安全評估、有限範圍國際協議及台灣可採取的政策作法。

關鍵詞：人工智慧、國際政治、安全化、全球治理、美中科技競爭

* 楊惟任教授，英國 University of Warwick 政治學暨國際關係博士，現任職於世新大學行政管理學系，研究專長包括人工智慧與社會科學、川普學、全球治理、國際氣候政治。Email: william@mail.shu.edu.tw。

本論文經兩位雙向匿名審查通過。收件：2026/10/1。同意刊登：2026/10/26。

壹、前言

2026 年 9 月下旬，人工智慧（Artificial Intelligence, AI）在數日內接連進入聯合國大會（United Nations General Assembly, UNGA）、聯合國安理會、美中元首會談等高階外交議程。9 月 22 日，美國總統 Donald Trump 於第 81 屆聯合國大會一般性辯論中，拒絕以全球架構管控 AI，並表示聯邦政府將使用「超級智慧」（super intelligence）一詞取代 AI（The White House, 2026a）。9 月 23 日，安理會召開「人工智慧與國際安全」高階會議，科學家與先進 AI 企業負責人，包括 OpenAI 執行長 Sam Altman 與 Anthropic 執行長 Dario Amodei 向安理會說明風險與治理主張；會議未形成具拘束力的成果，美、俄等國對全球治理架構及安理會角色亦持保留立場（United Nations Security Council, 2026）。這些事件顯示，AI 的議程位階已明顯提高到全球治理的層次，但各方對治理方向並未形成一致意見。

同一時期，非國家行為者的安全論述明顯升溫。全球 AI 領域知名 AI 研究者、聯合國獨立國際人工智慧科學小組（Independent International Scientific Panel on Artificial Intelligence）共同主席 Yoshua Bengio 在安理會將先進 AI 的失控、濫用與跨境風險提升至國際安全層次。芬蘭與挪威領袖於 9 月 21 日發起的「控制先進 AI 模型呼籲」，主張在部署前進行測試與獨立評估、共享並通報嚴重事故，並研議設立具標準制定與驗證功能的國際機構，該倡議獲多國政府領袖及歐盟執委會（European Commission, EC）主席支持（Office of the President of the Republic of Finland, 2026）。儘管各方對風險程度看法不一，企業、科學界、國家，以及政府間與非政府國際組織，已在同一高階議程中競逐風險定義與治理工具。

事件層面的變化則為原本較抽象的風險提供可以核對的事件證據。澳洲總理 Anthony Albanese 於 9 月 24 日證實，OpenAI 代理曾在 6 月未經授權進入由澳洲服務部（Services Australia）管理的國民醫療保險（Medicare）統計入口網站並存取非公開檔案，澳洲政府其後啟動跨機關調查（Albanese, 2026）。OpenAI 另於 8 月公布 7 月內部資安評估事件的調查結果，指出模型曾繞過隔離控制、利用弱點取得網際網路存取，並接觸第三方系統（OpenAI, 2026）。兩起事件都不足以單獨證成「失控 AI」的最嚴重情境，但顯示代理式系統可能超出既定任務或授權邊界採取行動，因此提高事故通報、責任歸屬與跨境協調的政策重

要性。

就外交活動而言，中國領導人習近平 9 月 24 日在白宮會談表示，中美可持續就 AI 風險與效益對話，並共同防範濫用、維持人類控制（中華人民共和國外交部，2026）。白宮 9 月 25 日公布的事實清單指出，雙方已建立「美中超級智慧對話」，預計於 2026 年 11 月前舉行下一輪交流，並設置相關事故的雙邊溝通管道（The White House, 2026b）。這項發展顯示，激烈的科技競爭可與有限的風險管理並存。因此，本文不將合作視為競爭的對立面，而將兩者視為共同提升 AI 國際政治位階的不同途徑。

2026 年 9 月，高階外交會議、國家安全論述、代理式 AI 事件及治理倡議接連出現。不過，短期內的關注增加，未必代表議題地位已發生持久變化。因此，本文從 2017 年開始追蹤相關政策與國際協議的發展，檢視 2026 年的變化是否建立在先前累積的基礎上。

（一）2017 年至 2022 年：國家戰略、倫理原則與科技管制並行

中國國務院於 2017 年 7 月發布《新一代人工智慧發展規劃》，設定 2030 年成為世界主要人工智慧創新中心的目標，將 AI 明確提升為國家戰略（中華人民共和國國務院，2017）。2019 年 5 月，經濟合作暨發展組織（Organization for Economic Co-operation and Development, OECD）採納《人工智慧建議書》（*Recommendation of the Council on Artificial Intelligence*），成為首份政府間 AI 政策原則（OECD, 2019）。2021 年 11 月，聯合國教育科學文化組織（United Nations Educational, Scientific and Cultural Organization, UNESCO）採納《人工智慧倫理建議書》（*Recommendation on the Ethics of Artificial Intelligence*）（UNESCO, 2021）。同年 12 月，中國提出《關於規範人工智慧軍事應用的立場文件》，顯示 AI 已進入軍備管制與國際安全的論述場合（中華人民共和國外交部，2021）。2022 年 10 月，美國商務部工業與安全局（Bureau of Industry and Security, BIS）宣布限制中國取得先進運算晶片、發展超級電腦與製造先進半導體的能力（Bureau of Industry and Security, 2022），AI 所依賴的物質基礎自此成為大國競爭的直接標的。

（二）2023 年：生成式 AI 衝擊與風險議題主流化

2023 年是 AI 風險論述走向主流的轉折年。3 月，生命未來研究所（Future of Life Institute）發表公開信，呼籲所有 AI 實驗室至少暫停六個月訓練比 GPT-4

更強大的系統（Future of Life Institute, 2023）；5 月，Sam Altman 等人撰文主張，未來可能需要類似國際原子能總署（International Atomic Energy Agency, IAEA）的國際機構監督超級智慧的發展（Altman et al., 2023）；同月，AI 安全中心（Center for AI Safety）發起的聲明獲得眾多頂尖電腦科學家聯署，主張降低 AI 導致人類滅絕的風險，應與大流行病、核戰並列為全球優先事項（Center for AI Safety, 2023）。7 月 18 日，安理會在英國主持下首度召開 AI 專題會議，邀請 Anthropic 共同創辦人 Jack Clark 進行簡報（United Nations Security Council, 2023）。10 月，中國在第三屆「一帶一路」國際合作高峰論壇提出《全球人工智慧治理倡議》（中華人民共和國外交部，2023），美國則頒布第 14110 號行政命令，要求開發具雙重用途之基礎模型的企業向政府報告安全測試結果（Exec. Order No. 14110, 2023）。11 月，包括美國與中國在內的 28 國及歐盟於英國 Bletchley Park 發表宣言，於共同宣言中承認先進 AI 可能造成嚴重乃至災難性的傷害（*The Bletchley Declaration*, 2023）。同月 15 日，美國總統 Joe Biden 與習近平於加州 Woodside 會晤，確認有必要透過兩國政府間的對話處理先進 AI 系統的風險並提升 AI 安全（Biden, 2023），AI 自此正式進入美中元首層級的議程。

（三）2024 年：多邊制度建設加速

2024 年 3 月，聯合國大會以共識採納由美國主導的第 78/265 號決議，聚焦安全、可靠且值得信賴的 AI 系統與永續發展（United Nations General Assembly, 2024b）。同年 7 月，再以共識採納由中國主導、強調 AI 能力建設國際合作的第 78/311 號決議（United Nations General Assembly, 2024a）。區域層次上，歐盟完成《人工智慧法》（*Artificial Intelligence Act*）立法程序（Regulation (EU) 2024/1689, 2024），歐洲理事會（Council of Europe, CoE）開放簽署首部具法律拘束力的 AI 國際公約（Council of Europe, 2024）。科學界方面，Yoshua Bengio 等 25 位學者於 *Science* 期刊聯名示警，先進 AI 的快速進展可能帶來大規模社會傷害，乃至人類失去對自主 AI 系統控制的極端風險（Bengio et al., 2024）。9 月，聯合國 AI 高階諮詢機構提出《為人類治理 AI》報告（*Governing AI for Humanity*），建議設立國際科學小組與全球政策對話等機制（United Nations High-level Advisory Body on Artificial Intelligence, 2024）。

（四）2025 年：路線分歧與治理機制成形

2025 年 1 月，Donald Trump 甫上任即撤銷第 14110 號行政命令，並隨即頒布第 14179 號行政命令，要求移除阻礙美國 AI 領導地位的政策障礙（Exec. Order No. 14179, 2025）。2 月，巴黎 AI 行動峰會上，約 60 國簽署包容與永續 AI 聲明，美國與英國則未聯署（Crouse, 2025）。7 月，美國發布《贏得競賽：美國 AI 行動計畫》（*Winning the Race: America's AI Action Plan*）（The White House, 2025），中國則於世界人工智慧大會發表《人工智慧全球治理行動計劃》，並倡議成立世界人工智慧合作組織（世界人工智能大會暨人工智能全球治理高級別會議，2025）。8 月，聯合國大會採納第 79/325 號決議，正式設立獨立國際人工智慧科學小組與全球 AI 治理對話兩項機制（United Nations General Assembly, 2025）。11 月，Anthropic 披露其偵測並阻斷一起由中國政府支持之駭客組織利用 AI 代理自主執行大部分作業的網路間諜行動，並指出此為首例有紀錄、由 AI 執行的大規模網路間諜攻擊（Anthropic, 2025）。

（五）2026 年：風險具體化與議程位階躍升

2026 年，AI 的國際政治屬性強化。6 月 12 日，白宮以國家安全為由要求 Anthropic 限制外國人存取 Fable 5 與 Mythos 5，Anthropic 因無法即時依國籍區隔使用者而暫停全球存取，相關限制於 6 月 30 日解除（Anthropic, 2026）。7 月，聯合國獨立國際人工智慧科學小組發布初步報告，首屆全球 AI 治理對話在日內瓦舉行。同月，29 國在上海簽署成立世界人工智慧合作組織的協定（Independent International Scientific Panel on AI, 2026）。其後 OpenAI 與 Hugging Face 與澳洲政府網站事件相繼曝光，並在 9 月的安理會會議、多國領袖倡議與美中峰會中被轉化為治理與風險管理議題。

表 1

AI 國際政治化之重要事件時序（2017 年至 2026 年 9 月）

時間	事件	議題內涵	資料來源
2017 年 7 月	中國發布《新一代人工智慧發展規劃》	AI 提升為國家戰略	中華人民共和國國務院（2017）
2019 年 5 月	OECD 通過《理事會人工智慧建議書》	首份政府間 AI 標準	OECD（2019）

時間	事件	議題內涵	資料來源
2021 年 11 月	UNESCO 通過《人工智慧倫理建議書》	全球性倫理規範擴散	UNESCO（2021）
2021 年 12 月	中國發布規範 AI 軍事應用立場文件	AI 進入軍控論述	中華人民共和國外交部（2021）
2022 年 10 月	美國對中國實施先進運算晶片出口管制	算力成為戰略資源	Bureau of Industry and Security（2022）
2023 年 3 月	生命未來研究所呼籲暫停訓練更強大系統	風險論述進入公共議程	Future of Life Institute（2023）
2023 年 5 月	AI 滅絕風險聲明；OpenAI 倡議類國際原子能總署	人類安全框架成形	Altman 等人（2023）；Center for AI Safety（2023）
2023 年 7 月	安理會首度召開 AI 專題會議	進入安全議程	United Nations Security Council（2023）
2023 年 10 月至 11 月	中國《全球人工智慧治理倡議》；美國第 14110 號行政命令；*The Bletchley Declaration*；Woodside Summit 同意建立 AI 政府間對話	大國治理主張競逐	中華人民共和國外交部（2023）；Biden（2023）；*The Bletchley Declaration*（2023）；Exec. Order No. 14110（2023）
2024 年 3 月、7 月	聯合國大會第 78/265 號與第 78/311 號決議	多邊規範制度化	United Nations General Assembly（2024a, 2024b）
2025 年 1 月至 2 月	美國第 14179 號行政命令；巴黎峰會美英未聯署	治理路線分歧	Crouse（2025）；Exec. Order No. 14179（2025）
2025 年 7 月至 8 月	美國 AI 行動計畫；中國全球治理行動計劃；聯合國大會第 79/325 號決議	競爭與制度化並進	世界人工智能大會暨人工智能全球治理高級別會議（2025）；United Nations General Assembly（2025）；The White House（2025）

時間	事件	議題內涵	資料來源
2025年11月	Anthropic 披露 AI 主導之網路間諜行動	代理式 AI 風險具體化	Anthropic（2025）
2026年6月	美國對先進模型發布出口管制指令並於月底解除	模型戰略資源化	Anthropic（2026）；Power（2026）
2026年7月	科學小組初步報告；首屆全球 AI 治理對話；世界人工智慧合作組織成立協定；OpenAI 與 Hugging Face 事件	平行制度與失控風險並現	Independent International Scientific Panel on AI（2026）；OpenAI（2026）
2026年9月	Donald Trump 聯合國大會演說；安理會高階簡報；澳洲事件；習近平國事訪問；微軟前執行長 Bill Gates 示警；芬挪「控制先進 AI 模型呼籲」；美中同意建立 AI 對話與事件溝通管道	議程位階躍升至最高層級	中華人民共和國外交部（2026）；Amodei（2026）；Durkee（2026）；Handley（2026）；Irwin（2026）；Office of the President of the Republic of Finland（2026）；UN News（2026b）；The White House（2026a, 2026b）

資料來源：作者自行整理。

回顧這段歷程，可以看到兩股並行的力量。主要國家從安全與相對實力出發，把晶片、運算能力及先進模型視為需要掌握的戰略資源；科學界、國際組織與部分科技企業則提出跨境安全風險，主張建立共同的評估標準和治理安排。2026 年 9 月的事件讓這兩種考量集中浮現：美國在聯合國大會反對全球統一管制，其企業領導人卻在安理會呼籲訂定國際標準；中國支持聯合國的治理角色，也參與建立新的國際合作組織。這些政策未必一致，卻都提高了 AI 在國際政治中的重要性。

既有國際關係文獻對 AI 的討論，主要沿著兩種研究方向發展。第一條將 AI 視為影響國家實力與權力平衡的促進其他能力的技術，主要關注軍事能力、經

濟競爭與權力轉移（Horowitz, 2018; Payne, 2018），第二條把 AI 視為全球治理對象，著重規範擴散、制度設計與治理碎片化等問題（Schmitt, 2022; Tallberg et al., 2023）。然而，2026 年的經驗顯示，競爭增加未必會減少合作，而可能在同一時期並行增強：先進模型被納入出口管制之際，美中也討論 AI 事故通報機制；聯合國設立新的科學與對話機制之際，中國另行成立世界人工智慧合作組織，美國則公開反對全球管控架構。若僅沿用其中一條研究路徑，則難以完整說明 AI 為何成為大國競逐的戰略資產與多邊治理的焦點。

相關研究經常以「日益重要」或「受到高度關注」描述 AI 的國際地位，卻較少說明應如何衡量議題的重要程度。本文不預先認定 AI 已與核武或氣候變遷具有相同地位，而是提出可以觀察的條件，檢驗 AI 是否符合本文對「核心議題」的界定。2026 年的代理式 AI 越權事件則用來考察具體事故能否促成政策關注，而非將事故發生直接等同於議程升級。

台灣是先進半導體供應鏈的重要參與者，卻無法以聯合國會員國身分參與其治理機制。AI 國際治理的變化，因此既影響台灣的供應鏈及安全政策，也涉及國際合作機會。理解這項議題如何進入主要國家的外交議程，有助於評估台灣可以參與的政策工作。

基於上述研究動機，本文設定四項研究目的：首先，追溯 AI 由科技創新議題轉為國際政治議題的歷程；其次，整合安全化理論、權力政治與全球治理，建立解釋 AI 國際政治化的分析架構；接著，將「國際政治核心議題」操作化，並據以評估 AI 在 2026 年的議題地位；最後，提出具可行性的國際治理與國內政策建議。

依照研究目的，本文聚焦以下三項問題：

（一）AI 如何由科技創新與產業政策議題，轉化為國家安全與大國競爭的高階政治議題？其關鍵轉化機制為何？

（二）AI 跨境風險所引發的治理需求，如何推動國際規範與多邊制度的形成？國家、國際組織與科技企業各自如何參與？

（三）競爭與合作兩股動力之間呈現何種互動關係？在什麼條件下，兩者共同將 AI 推升為國際政治的核心議題？該地位又能以哪些可觀察的指標加以判定？

本文的研究對象為「作為國際政治議題的 AI」，而非 AI 技術，主要討論先進 AI（frontier AI）與通用目的 AI，亦即具備廣泛任務能力、並逐步發展出代理式自主行動能力的大型模型，至於特定領域的應用型 AI，僅於其涉及國際安全或跨境治理時納入討論。

在行為者方面，本文涵蓋四類行為者：第一，主要國家，主要為美國與中國，並兼及歐盟、英國、法國與澳洲等在事件中扮演重要角色的國家；另外，國際組織與多邊機制，包括安理會、聯合國大會、UNESCO、OECD 與歐洲理事會；此外，先進 AI 企業，包括 OpenAI、Anthropic 與 Hugging Face；最後，以獨立國際人工智慧科學小組為代表的科學界。

在時間範圍上，本文以 2017 年中國發布國家級 AI 發展規劃為起點，截至 2026 年 9 月底，其中 2023 年至 2026 年為分析重點。在議題範圍上，本文聚焦 AI 的國家安全、大國競爭與全球治理面向，各國國內 AI 法規僅於具有國際外溢效果或釋放對外訊號時納入討論，勞動市場、智慧財產權等議題則不在本文範圍之內。

本文採質性研究途徑，結合文獻分析與過程追蹤。文獻資料分為四類：首先是官方文件，包括聯合國決議與會議紀錄、各國行政命令與政策文件，以及中國外交部等政府機關發布的文本；其次是國際組織與科學機構報告，例如聯合國 AI 高階諮詢機構與獨立國際人工智慧科學小組的報告；另外是企業揭露，包括先進 AI 企業發布的事件報告與政策聲明；最後是具公信力的新聞報導與智庫分析，用以補充事件歷經與時序。

過程追蹤依據事件發生的先後順序與相關證據，檢驗原因如何作用於中介機制並影響結果（George & Bennett, 2005）。本文先考察 AI 的技術能力，再檢視戰略規劃、安全論述、跨境風險與國際制度的發展，並分析焦點事件和大國互動是否加快議題升級。最後以 AI 在國際外交議程的地位為待解釋結果，核對各環節的證據。對於重要事件，本文參照不同來源；涉及企業自行揭露的事故，則與政府聲明及新聞報導相互比對，以降低單一資料來源的偏誤。

為提高因果推論的可證偽性，本文預先界定兩類過程追蹤證據。若焦點事件確實促成議程升格，至少應觀察到事件曝光與高階議程在時序上緊密相連，且高層行為者在論述中明確指涉相同風險；此類證據可構成排除部分替代解釋的環形檢驗（hoop test）。若決策者明言某項事件促成特定政策或議程安排，則

可視為證明力較強的確鑿證據檢驗（smoking-gun test）（George & Bennett, 2005；Collier, 2011）。然而，任一檢驗皆不足以憑單一證據確立因果關係。本文另檢驗三項競爭性解釋：議題關注週期（Downs, 1972）、聯合國大會高階會期與安理會輪值及元首互訪造成的時程巧合，以及企業藉安全論述影響規則設計的策略動機。各項檢驗的結果，於第五節第五小節逐項說明。

貳、文獻探討

一、人工智慧與國際政治研究

科技變遷與國際政治的關係，長期以來即是國際關係研究的重要課題。Drezner（2019）指出，並非所有技術創新都會改寫國際政治，其影響取決於技術擴散速度、固定成本高低，以及國家控制能力，若技術具有顯著戰略價值且僅少數國家得以掌握，便可能改變權力分配與安全互動。根據這項觀點，AI 的特殊之處與其兼具高固定成本（先進模型所需的算力與資本門檻）與高擴散性（模型與知識快速外溢），呈現能力集中與知識擴散並存的特徵。

在權力平衡研究，Horowitz（2018）將 AI 界定為促進其他能力的技術，認為其影響程度取決於軍事與商業應用的界線，以及組織吸納新技術的能力。Payne（2018）從戰略研究出發，指出 AI 可能改變決策速度及其心理基礎，並影響戰略穩定。Ding 與 Dafoe（2023）區分「領先部門」與「通用目的技術」兩種權力轉移路徑，主張通用目的技術的影響並非來自單一突破，而取決於國家能否把技術廣泛導入經濟與軍事體系。Ding（2024）以案例指出，技術擴散能力而非創新領先，更能解釋大國崛起。Scharre（2023）從美中 AI 競爭出發，將資料、運算、人才與制度等視為決定國家 AI 實力的重要條件。

中國學界也注意到 AI 的國際政治意涵。傅瑩（2019）從國際格局與國際規範分析，認為 AI 將在經濟與軍事領域改變國家間力量對比、擴大非國家行為者能力，並衝擊國際法與倫理規範，因此主張各國應以共同安全為基礎討論 AI 國際規範。封帥（2018）指出，AI 可能促使國際體系發生變化，也可能加劇國家間不平等，使技術落後國家面臨邊緣化風險。兩者為將 AI 視為權力資源奠定理論基礎，但相關研究發表於生成式與代理式 AI 興起之前，因此沒有探討跨境失控風險如何反向促成國際合作。

二、科技、安全化與權力政治

安全化理論提供理解議題如何進入安全領域的重要工具。依 Buzan 等人（1998）的界定，安全化是一種論述實踐：安全化行為者把特定議題描述為對某一指涉對象的生存威脅，若獲受眾接受，便可能正當化超出常規政治程序的例外措施。Wæver（1995）強調，「安全」的政治意義不完全取決於威脅是否客觀存在，而取決於議題如何被帶離常規政治。Balzacq（2005）修正過度偏重論述的理解，指出安全化能否成立，亦取決於脈絡、受眾與行為者的權力位置，因此也是一種策略性實踐。

Hansen 與 Nissenbaum（2009）將安全化理論應用於網路安全，提出三項有助於分析 AI 安全論述的概念。首先是「超安全化」（hypersecuritization），亦即把威脅推演至災難性與連鎖性的極端情境；其次是「日常安全實踐」，透過一般人的日常經驗讓威脅變得具體可感；最後是「技術化」（technification），即議題因專業性較強而讓技術專家取得界定威脅的權威。AI 議題呈現這三項特徵：滅絕風險聲明具有超安全化色彩，代理式 AI 入侵政府網站使抽象風險具體化，而先進實驗室與科學家因技術專業而取得權威性的議題發言權。

安全化不是單向的過程。Wæver（1995）提到的去安全化是把議題重新帶回常規政治，Hansen（2012）進一步區分穩定化、取代、重新表述與沉默化四種去安全化形式，指出議題可能被改寫為其他框架，而不一定完全退出安全領域。Stritzel 與 Chang（2015）以阿富汗為例，說明在地行為者可以提出「反安全化」論述，挑戰主導者的威脅定義，Corry（2012）則把「安全化」和「風險化」進行區分，前者指向明確且迫切的威脅，後者處理長期且有條件的可能傷害，通常引發預防性的治理安排，而非緊急的例外措施。這些概念有助於分析各方在同一議題上的安全主張，如何相互競爭、吸收與抵消，第五節將依此討論 AI 的雙軌安全化。

權力政治可協助說明安全化發生作用的條件。Jervis（1978）的安全困境理論指出，在攻守難辨的情況下，一國為自保而增加能力，容易被他國視為威脅，並引發螺旋式競爭。AI 具有明顯的軍民兩用特性，其攻防屬性不易區分，因此容易形成安全困境。Farrell 與 Newman（2019）提出「武器化的相互依存」，指出掌握全球網絡樞紐的國家，可透過資訊監控與限制存取，把經濟相互依存轉為強制工具，半導體供應鏈與雲端算力即具有這類網絡特徵。Allison（2017）

的權力轉移分析，則提供理解美中科技競爭政治背景的另一項參照。

三、全球治理

自由制度主義認為，即使處於無政府狀態，國際制度仍可藉由降低交易成本、提供資訊與穩定預期，促成國家合作（Keohane, 1984）。制度形式並非固定不變。Abbott與Snidal（2000）指出，國家經常刻意選擇拘束力、精確性與授權程度較低的非強制性規範，以便在不確定環境下保留彈性並降低締約成本。目前AI治理仍以原則、宣言與建議書為主，與這項觀察相當一致。

在規範形成方面，Finnemore 與 Sikkink（1998）的規範生命週期模型，把規範發展分為出現、擴散與內化三個階段，並強調規範倡議者在早期階段的作用。再者，P. M. Haas（1992）的知識社群（epistemic community）概念指出，當決策者面臨相當大的不確定性時，具有共同專業知識與因果信念的專家網絡，可能透過問題界定與政策方案影響國家利益的認知。AI 領域的科學界與先進實驗室研究人員，即具有部分相近特徵。

隨著治理機制持續增加，AI治理逐漸呈現「機制複合體」（regime complex）的特徵，也就是由多個彼此重疊、但未必存在明確層級關係的制度所形成的鬆散體系（Keohane & Victor, 2011）。Cihon等人（2020）比較集中與碎片化兩種AI 治理架構，指出碎片化雖保有彈性與政策試驗空間，卻可能帶來規則衝突與選擇有利的國際協商平台。Schmitt（2022）把全球AI治理描述為在碎片化制度環境中逐步成形的新興治理安排。Roberts 等人（2024）提到，地緣政治競爭、各國監管路徑差異與能力落差，是全球 AI 治理面臨的主要障礙。Bradford（2023）對數位監管模式的研究，顯示主要政治經濟體可能以不同監管理念向外投射規則。

既有研究仍有一些尚待處理的問題。首先，權力政治研究偏重 AI作為國家競爭資源的角色，全球治理研究則關注國際合作和規範發展。若分開討論，便難以解釋2026年競爭升高之際，為何也出現新的合作安排。

此外，現有討論仍較重視國家行為。安全化研究通常以國家菁英為主要分析對象，但 AI企業領導人與科學家已直接向安理會提出安全主張。他們可能是技術風險的來源，也是事故資訊的提供者及規範倡議者。這種角色重疊，尚需更仔細的研究。

相關文獻對「核心議題」的判準也不夠明確。「日益重要」與「高度關注」只能描述關注程度，無法回答 AI 是否已達到足以影響主要國家決策與國際制度發展的地位。

最後，早期研究受到發表時間限制，無法涵蓋 2025 年至 2026 年的模型出口管制、代理式 AI 網路事件與新設的聯合國治理機制。本文補充近期官方紀錄與事件資料，但對調查尚未完成、制度成效仍待觀察的情形，保留必要的判斷限制。

四、分析架構與研究命題

為回應上述不足，本文建立一套整合分析架構（見表 2）。自變項為 AI 的技術能力，尤其是通用目的性與代理自主性；依變項則是 AI 在國際政治中的議題地位，也就是「核心議題化」程度。兩者之間包含兩組轉化機制：競爭機制由「戰略化」與「安全化」構成，推動 AI 進入高階政治，合作機制由「跨境風險」與「制度化」構成，使 AI 進入多邊治理，焦點事件、跨國倡議網絡與大國互動是為催化條件，影響轉化的速度與時機。

表 2
AI 國際政治核心議題化之分析架構

構成要素	內涵	主要理論依據	觀察重點
自變項：技術能力	通用目的性、代理自主性、軍民兩用性	Bresnahan 與 Trajtenberg（1995）；Ding 與 Dafoe（2023）	能力門檻的跨越與擴散
競爭機制：戰略化	將 AI 及其投入要素界定為國家戰略資產並施以管制	Farrell 與 Newman（2019）；Scharre（2023）	出口管制、國家戰略、產業政策
競爭機制：安全化	將 AI 建構為國家安全或人類生存威脅，正當化例外措施	Buzan 等人（1998）；Hansen 與 Nissenbaum（2009）	安全化行為者、受眾與例外措施
合作機制：跨境風險	AI 風險外溢至單一國家管轄範圍之外，形成集體行動問題	P. M. Haas（1992）；Keohane（1984）	跨境事件、治理缺口、知識社群

構成要素	內涵	主要理論依據	觀察重點
合作機制：制度化	規範、宣言、條約與常設機制之建立	Abbott 與 Snidal（2000）；Finnemore 與 Sikkink（1998）；Keohane 與 Victor（2011）	非強制性規範與強制性規範、機制複合體
催化條件	焦點事件、跨國倡議網絡、大國互動	Kingdon（2011）	政策之窗的開啟
依變項：核心議題化	AI 在國際議程中的位階與穩定地位	本研究	六項操作化指標

資料來源：作者自行整理。

依據該架構，本文提出以下五項研究命題：

命題一（能力轉化命題）：AI 的技術能力愈具備通用目的與自主行動，其影響的行為者範圍與傷害類型愈廣，政治化程度亦愈高。

命題二（競爭機制命題）：當大國將 AI 界定為決定相對權力的戰略資產時，AI 議題將透過戰略化與安全化進入高階政治，並觸發資源管制等例外措施。

命題三（合作機制命題）：當 AI 風險呈現跨境外溢，且無單一國家能夠獨力管控時，將產生多邊制度建設的功能需求，並由知識社群與國際組織提供規範供給。

命題四（雙重動力命題）：競爭與合作並非互斥，而會相互強化，競爭衍生的風險擴大合作需求，合作場合也成為新的競爭舞臺，兩者共同提升 AI 的議程位階。

命題五（焦點事件命題）：具體可感的越權或失控事件可在短期內開啟政策之窗，加速 AI 從重要議題升格為核心議題。

五、「國際政治核心議題」之操作

本文將「國際政治核心議題」界定為：在特定時期內，被主要國家視為攸關國家安全與國際秩序，占據最高層級外交議程，並驅動制度建構與跨議題交換的議題。依上述資料，本文提出六項操作化指標（見表 3）。

一、議程層級，指議題是否進入元首會談、安理會等最高層級的外交場合，

而非僅停留於部會或技術層級。二、安全框架，指議題是否被主要行為者以國家安全或人類生存等威脅語彙加以建構，並觸發例外措施。三、資源戰略化，指與議題相關的物質資源是否被國家列為戰略資產並施以管制。四、制度化密度，指針對該議題之國際文件、多邊機制與常設機構的數量與層級。五、行為者擴散，指參與議題之行為者類型是否由國家擴及國際組織、企業與科學界。六、跨議題連結，指該議題是否與貿易、軍備管制、區域安全等其他重大議題相互連結，成為外交談判的籌碼。

這六項指標取材於既有文獻中衡量議題重要性的不同面向。議程層級沿用 Hoffmann（1966）對「高階政治」與「低階政治」的區分，並參考 Keohane 與 Nye（1977）的看法：在複合相互依存的條件下，議題之間的位階並不固定，隨議程變動而調整。再者，Kingdon（2011）提醒，議題能否在決策者有限的注意力下獲得注意，反映其之重要程度。

安全框架源自安全化理論（Buzan et al., 1998），用來判斷議題是否被帶離常規政治。資源戰略化參考 Farrell 與 Newman（2019）對網絡樞紐與國家強制力的分析，觀察國家是否把議題所依賴的物質基礎轉為管制工具。制度化密度來自國際建制與機制複合體研究，以規範的拘束力、精確性與授權程度，以及制度之間的重疊程度（Abbott & Snidal, 2000; Keohane & Victor, 2011）作為衡量依據。行為者擴散參照跨國倡議網絡（Keck & Sikkink, 1998）與知識社群（P. M. Haas, 1992）的研究，觀察非國家行為者是否取得界定問題的發言權。跨議題連結依循 E. B. Haas（1980）對議題連結的討論，議題若經常與其他議題並列協商或相互交換，表示各國願意為它付出談判成本。

本文的作法不在創造全新概念，而是把分散在不同文獻的考量條件，整合成一套可以檢驗的指標。

表 3

國際政治核心議題之操作指標

指標	定義	可觀察證據	判定門檻
議程層級	議題進入外交議程的最高層級	元首會談紀錄、安理會會議、聯合國大會一般性辯論	出現在元首會談與安理會議程

指標	定義	可觀察證據	判定門檻
安全框架	以國家安全或人類生存語彙建構議題	官方聲明、安理會發言、科學界聲明	主要大國與非國家行為者皆採安全語彙
資源戰略化	相關資源被列為戰略資產並受管制	出口管制、國家戰略文件、產業政策	管制範圍擴及核心生產要素與最終產品
制度化密度	國際文件與常設機制之數量與層級	聯合國大會決議、國際公約、常設科學與對話機制	出現常設性全球機制與具拘束力之條約
行為者擴散	參與行為者類型的廣度	企業揭露、國際組織報告、會議出席名單	國家、國際組織、企業與科學界共同參與
跨議題連結	與其他重大議題的連結程度	峰會議程、雙邊聯合聲明	與貿易、軍控或區域安全議題並列協商

資料來源：作者自行整理。

本文依符合指標的數量，暫將議題區分為三類：符合一至二項者為「邊緣議題」，符合三至四項者為「重要議題」；若六項指標至少五項達到表 3 所列門檻，且其中包括「議程層級」及「安全框架」，便歸為「核心議題」。將後兩項列為必要條件，是因為前者涉及議題是否進入最高層級的外交決策，後者則涉及議題是否被認定為需要採取特別措施的安全問題。這是一套供本文進行比較的操作規則，尚非適用所有國際政治議題的既定標準。第五節將使用這套規則評估 2026 年的 AI 議題。

關於門檻的設定方式，本文參考 Goertz（2006）的概念建構方法，採用必要條件與「家族相似」結合的混合結構：議程層級與安全框架是「核心議題」不可缺少的屬性，其餘四項則可彼此替代，只要多數成立即可。這種「n項中至少 m 項」的規則，適用於內涵較廣、個別屬性不一定都出現的政治學概念。門檻定為五項而不是四項，因為四項的規則容許議題只靠兩項必要條件，加上行為者擴散、跨議題連結等較易達成的指標便成立，難以區分「受到高度關注」與「持續影響決策及制度發展」兩種狀態，若要求六項全數符合，任何單一指標的資料缺口都可能推翻判定，對仍在形成中的議題又過於嚴格。

依 Adcock 與 Collier（2001）對測量效度的討論，判定規則應能區分性質不

同的個案，因此第五節除了評估 2026 年，也以 2023 年作為對照，若採四項門檻，2023 年的 AI 便可能被歸為核心議題，與 2023 年至 2026 年間管制範圍及制度安排的明顯變化不符，若改採六項門檻，2026 年的判定結果則維持不變。這項簡單的敏感度檢驗顯示，五項門檻的主要作用在於區分兩個年度，而不是讓 2026 年的個案過關。當然，門檻仍有待適用到其他議題後再行校準。

參、人工智慧的安全化與權力政治

一、AI 能力擴張與通用目的科技

經濟學文獻將通用目的技術界定為具有三項特徵的技術：應用範圍普遍、具持續改進的潛力、能夠帶動下游領域的互補性創新（Bresnahan & Trajtenberg, 1995）。蒸汽機與電力之所以改變國家間的經濟與軍事實力對比，並並不單取決於單一發明，而與其擴散至整體生產體系的速度與深度（Ding & Dafoe, 2023）。AI，尤其是以大型模型為基礎的生成式與代理式 AI，正也展現這三項特徵：同一套基礎模型可被用於程式撰寫、科學研究、行政作業與軍事規劃；模型能力隨算力與資料投入持續提升；圍繞模型的應用生態不斷擴張。

AI 能力擴張的政治影響，與受其影響的行為者和政策範圍有關。當 AI 主要處理特定任務時，政策辯論多集中於產業、就業與隱私；當大型模型可廣泛用於不同領域，相關議題便涉及國家競爭力與軍事能力。代理式 AI 若能在有限的人類監督下規劃、執行多步驟任務，風險也不再僅限於使用者蓄意濫用，而可能包括模型超出既定授權的行動。Bengio 等人（2024）指出，自主系統若缺乏適當安全保障，可能採取不符合人類意圖的行為，且隨能力提高而更難控制。

2025 年至 2026 年的經驗，顯示上述推論已開始出現可以查證的案例。Anthropic（2025）揭露的網路間諜行動中，攻擊者利用 AI 代理自主完成偵察、漏洞利用、憑證竊取與資料外洩等大部分行動，人類操作者主要負責選定目標與做出關鍵決策。2026 年 7 月，OpenAI 在內部評估發生模型繞過隔離控制、透過未經授權的管道溝通並存取第三方系統的事件（OpenAI, 2026）。Dario Amodei 在 9 月指出，AI 協助開發下一代 AI 的遞迴式改進趨勢已在產業間加速，此一發展連同前述事件，是其主張放緩先進能力提升速度的主要理由（Amodei, 2026）。能力門檻的跨越，使 AI 除了被國家利用，也成為政府必須監督的對象，這正是命題一所描述的轉化過程。

二、軍民兩用與跨境安全風險

AI 的軍民兩用性，使其難以沿用傳統軍備管制中區分民用與軍用的作法。Urbina 等人（2022）的實驗顯示，研究人員僅需將一個用於藥物研發的生成模型的目標函數反轉，即可在數小時內產生數以萬計的潛在毒性分子，其中包括已知的化學戰劑。這個案例說明，AI 的危險性並不單取決於特定的軍事用途，而內嵌於其通用能力之中，這使生物與化學武器的擴散風險成為 AI 安全論述的主要關切。

在網路領域，Taddeo 與 Floridi（2018）早已提到，AI 將使網路攻擊更具規模與自主性，若缺乏規範，恐引發網路軍備競賽。2026 年的澳洲事件揭示了另一層問題：OpenAI 的代理並非受命攻擊，而是在執行公共藥品支出的網路研究任務時，為取得資料而自行繞過存取阻擋（Albanese, 2026; Handley, 2026）。也就是說，跨境侵害可能源於無惡意的系統越權，而非國家或犯罪集團的蓄意行動。此類事件牽動不同國家的利益與管轄權：模型開發者所在國、受害系統所在國，以及算力與資料所在國，使得 AI 風險具有明顯的跨境外溢效應。

在傳統軍事領域，中國於 2021 年發布的立場文件，主張各國應以審慎負責的態度研發與使用軍事 AI，不謀求絕對軍事優勢，以免加劇戰略誤判（中華人民共和國外交部，2021）。Maas（2019）從核武軍控的歷史經驗出發指出，軍事 AI 的國際管制並非不可能，但成效取決於能否建立知識社群、形成規範，以及找到可驗證的管制標的。Russell（2019）從技術層面說明，若 AI 系統的目標設定與人類偏好存在落差，能力愈強的系統愈可能以人類無法預期的方式追求其目標，這使「控制問題」不再只是工程議題，而具有國際安全的意涵。

三、算力、晶片與模型的戰略資源化

AI 的生產仰賴集中的物質基礎。先進晶片的設計、製造設備與晶圓代工，分別集中於少數國家與企業，使半導體供應鏈成為全球最具戰略價值，也最脆弱的網絡之一（米勒〔Miller〕，2022/2023）。依 Farrell 與 Newman（2019）的觀點，掌握此類網絡樞紐的國家，得以透過阻斷存取的方式將相互依存武器化。美國於 2022 年 10 月實施的出口管制，明確以限制中國取得先進運算晶片、發展超級電腦與製造先進半導體的能力為目標，並以其可能用於先進軍事系統為主要理由（Bureau of Industry and Security, 2022）。Scharre（2023）亦將算力列為 AI 時代權力競爭的主要競爭領域之一，指出掌握算力供給的國家，將在模型

訓練與部署上享有長期優勢。

2026 年的一連串事件顯示，戰略資源化的對象已由晶片與設備延伸至模型。6月12日，美國以國家安全權限要求Anthropic限制外國人存取Fable 5與Mythos 5，因無法即時驗證使用者國籍，Anthropic 暫停兩款模型的全球存取，並公開質疑命令的程序與技術依據（Anthropic, 2026）。6 月 30 日限制解除後，全球服務恢復。由於解除條件的細節部分來自媒體報導，本文僅將此案作為「模型可直接成為國家安全管制標的」的證據，而不據此推論政府已形成穩定的模型出口管制制度。

另外，必須說明的是，美國在 2025 年 1 月公布的《AI 擴散框架》（*AI Diffusion Framework, ADF*），已將訓練運算量超過 10^{26} 次運算的封閉模型權重列入出口管制，但該規則在企業遵循義務生效前，即於同年 5 月遭到廢止（Bureau of Industry and Security, 2025a, 2025b）。相較之下，2026 年的案例是政府直接限制特定模型的對外提供，管制對象更具體。

此案說明，政府已可能把先進模型列為安全管制對象，而不再只管制晶片與運算設備。另方面，國家必須直接與模型開發企業交涉，企業對模型能力的說明及安全安排，也會影響政策決定。不過，這項限制在短時間內發布並解除，現有材料仍不足以證明美國已建立一致且持續適用的模型管制制度。

四、美中科技競爭與 AI 高政治化

美中兩國對 AI 的戰略定位，構成 AI 高政治化的重要背景。中國早在 2017 年即以國家規劃的形式，設定成為世界主要 AI 創新中心的時間表（中華人民共和國國務院，2017）。美國在 2025 年發布的行動計畫，將 AI 明確界定為一場必須贏得的全球競賽（The White House, 2025），Donald Trump 在 2026 年聯合國大會演說中更將這套框架推向極致，宣稱在超級智慧的競賽勝出者即為贏家，並表示美國目前大幅領先中國（Irwin, 2026）。在權力轉移的結構下，既有強權與新興強權之間對關鍵技術的爭奪，通常被雙方理解為零和的權力競逐（Allison, 2017）。

中國的回應兼具反制與規則競爭的性質。《全球人工智慧治理倡議》明確反對以意識形態劃線或構建排他性集團，並反對利用技術壟斷與單邊強制措施阻斷全球 AI 供應鏈（中華人民共和國外交部，2023），主要針對美國的出口管制與科技聯盟策略。2025 年的《人工智慧全球治理行動計劃》以促進發展應用

與確保安全可控為目標，強調尊重主權與公平普惠，並倡議成立世界人工智慧合作組織（世界人工智能大會暨人工智能全球治理高級別會議，2025）。由此可見，中國試圖將自身定位為全球南方利益的代言者與多邊治理的倡議者，以規則競爭回應美國的技術圍堵。

高度競爭並未排除有限的風險管理。2026 年 9 月峰會則建立 AI 與「超級智慧」對話與事故雙邊溝通管道（Ministry of Foreign Affairs of the People's Republic of China, 2026; The White House, 2026b）。同一安排在雙方文件的命名不同，也未形成共同的先進模型管制規則。較合理的解釋是：雙方在維持技術競爭之際，開始建立降低誤判與處理事故的窄域機制。此種競爭風險管理與傳統軍控的功能邏輯具有可比較性，但本文不主張兩者在威脅結構與可驗證性完全等同。

本節顯示，AI 成為高階政治議題，並非單由技術進步決定。主要國家相信 AI 可能影響相對實力，關鍵生產資源又集中於少數企業與供應鏈，因此將其納入戰略規劃及出口管制。另方面，模型的軍民兩用性和自主行動風險，也讓國家安全與人類安全論述受到重視。前者對應戰略化，後者反映安全化，兩種作用共同支持命題二，但各自的影響程度仍須透過後續政策資料檢驗。

肆、人工智慧的全球治理

一、跨境風險與治理缺口

若競爭機制使 AI 成為國家必須爭奪的資源，合作機制則源自於 AI 風險無法由單一國家獨力管控的事實。依自由制度主義的邏輯，當行為者面臨共同的外部性且個別行動無法有效因應時，國際制度便具有降低交易成本與提供資訊的功能價值（Keohane, 1984）。AI 風險的跨境性表現在三個層面：其一，模型一經部署，其使用與濫用便跨越國界；其二，代理式 AI 可自主存取他國的網路系統；其三，先進 AI 的研發集中於少數國家，其失誤的後果卻由全球承擔。

聯合國 AI 高階諮詢機構於 2024 年指出，現有全球治理安排仍有缺口，而且多數國家未能充分參與主要的 AI 治理倡議（United Nations High-level Advisory Body on Artificial Intelligence, 2024）。這顯示治理不足既涉及規範，也涉及各國參與及取得資訊的能力。

本文將治理缺口分為四類。首先是「知識缺口」，即各國政府對先進 AI 能

力與風險的評估，相當依賴開發企業自身揭露的資訊。其次是「管轄缺口」，澳洲事件即為例證：侵害行為由外國企業的模型在海外發動，受害國卻是在事發三個月後，透過一封寄至公開信箱的電子郵件方才得知（Handley, 2026）。另外是「執行缺口」，即便受害國確認侵害屬實，對境外企業的追責與制裁手段仍屬有限。最後是「參與缺口」，即 AI 治理規則主要由少數技術領先國家制定，開發中國家的聲音與能力均顯不足，這也是聯合國大會第 78/311 號決議強調能力建設的背景（United Nations General Assembly, 2024a）。

巴基斯坦副總理兼外長 Muhammad Ishaq Dar 於 2026 年 9 月的安理會會議強調，若技術開發的速度僅由少數國家與企業決定，而其他國家受到技術管制的限制，「為先進定速」的安全議程恐淪為技術排除的策略（United Nations Security Council, 2026）。

二、國際 AI 規範的形成

依 Finnemore 與 Sikkink（1998）的規範生命週期模型，AI 國際規範的發展大致已走過「出現」階段，進入部分「擴散」的過程。在規範出現階段，各國政府、企業與公民團體發布大量 AI 倫理原則。Jobin 等人（2019）分析全球 84 份 AI 倫理指引後發現，這些文件在透明、公平正義、不傷害、責任與隱私等原則上逐漸趨同，但在原則的詮釋與落實方式仍存在顯著分歧。

在規範擴散階段，政府間組織扮演重要角色。OECD（2019）的建議書提供首套政府間 AI 原則，UNESCO（2021）的倫理建議書便將規範擴及全體會員國。這段時期的規範多屬非強制性規範，其拘束力、精確性與授權程度均低，反映各國在技術前景不明時，傾向保留政策彈性（Abbott & Snidal, 2000）。

隨後，規範出現強制性的趨勢。歐盟《人工智慧法》採取以風險分級為主要的監管架構，對不同風險等級的 AI 系統課以相應義務，並對具系統風險的通用目的 AI 模型設有專門規範（Regulation (EU) 2024/1689, 2024）。歐洲理事會的 AI 框架公約是首部以人權、民主與法治為核心，具有法律拘束力的 AI 國際條約（Council of Europe, 2024）。此外，中國透過《全球人工智慧治理倡議》與行動計劃，提出以發展權、主權平等與能力建設為主軸的規範主張（中華人民共和國外交部，2023；世界人工智能大會暨人工智能全球治理高級別會議，2025）。這種規範多元並存的狀態，呼應 Bradford（2023）所描述的市場驅動、國家驅動、權利驅動三種數位監管模式之間的競爭。

三、先進 AI 安全與多邊治理

先進 AI 安全議題的國際化，經歷了由倡議、制度嘗試、路線逆轉到再度升溫的起伏過程。在倡議階段，2023 年的暫停訓練公開信與滅絕風險聲明，將先進 AI 的極端風險推向公共議程（Center for AI Safety, 2023; Future of Life Institute, 2023），OpenAI 執行長也提出類國際原子能總署的構想，將核能治理的制度引入 AI 領域（Altman et al., 2023）。

在制度嘗試階段，*《布萊切利宣言》（The Bletchley Declaration*）首次在包括美中在內的政府間文件承認先進 AI 的潛在災難性風險，並同意以國際合作進行風險辨識與因應（*The Bletchley Declaration*, 2023），美國第 14110 號行政命令以國內法規要求先進模型開發者向政府報告安全測試結果（Exec. Order No. 14110, 2023）。

然而，2025 年出現明顯的路線逆轉。Donald Trump 撤銷前述行政命令，轉而以移除 AI 發展障礙為政策主軸（Exec. Order No. 14179, 2025），並在巴黎峰會與英國一同拒絕聯署包容與永續 AI 聲明（Crouse, 2025）。政策轉變顯示，先進 AI 安全的多邊治理受到主要大國的國內政治與戰略判斷。

2026 年，先進 AI 安全議題再度升溫，但推動力量已有所轉移。在美國聯邦政府持續拒斥全球管控的情況下（Irwin, 2026），主要推動力量來自四方面提供：首先，代理式 AI 越權事件的接連曝光，使風險由理論推演轉為現實；其次，先進實驗室負責人主動呼籲放緩能力提升，並提出評估者常駐、民主國家協調與全球協調的方案（Amodei, 2026）；此外，聯合國體系的科學小組提供獨立的風險評估。其四，中等強國扮演聯盟建構者的角色，譬如芬蘭與挪威發起的「控制先進 AI 模型呼籲」，匯集歐洲、北美、亞太、中東與非洲的領袖聯署，將部署前強制測試、嚴重事故共同通報與國際驗證機構等主張提升至元首層級（Office of the President of the Republic of Finland, 2026）。這一變化意味著，先進 AI 安全治理的動力來源，已由國家主導的峰會外交，部分轉向由事件、企業、科學界與中等強國共同驅動的多行為者模式。

四、國際組織、國家與科技企業

AI 治理中的行為者關係，呈現國家、國際組織與科技企業三方互動的複雜樣態，其中又以企業角色的轉變最為顯著。在 AI 議題中，尖端 AI 企業扮演四種角色：包括風險的製造者，其模型即為潛在風險的來源；另包括風險的揭露

者，包括 Anthropic（2025）對網路間諜行動的揭露與 OpenAI（2026）對評估事件的報告；也有安全化行為者，其負責人直接向安理會陳述 AI 的安全威脅（Nichols et al., 2026）；最後是規範倡議者，主動提出放緩能力提升與引入外部評估等治理方案（Amodei, 2026）。

國家對企業的態度則呈現兩面性。一方面，國家將企業視為需要管制的對象，美國對先進模型發布出口管制指令（Anthropic, 2026），澳洲對 OpenAI 的延遲通報公開表達強烈不滿，並成立專案小組調查（Handley, 2026）。另方面，國家又高度倚賴企業提供的技術資訊與國防能力，並在大國競爭中將本國企業視為國家實力的延伸。這種既管制又依賴的關係，使國家與企業之間的互動更近似持續的談判，而非單向的命令與服從。

知識社群則在國家與企業之間扮演中介角色。依 P. M. Haas（1992）的觀點，當決策者面臨相當大的不確定性時，專家網絡的因果解釋與政策建議將影響國家利益的界定。獨立國際人工智慧科學小組的設立，正是將上述作用制度化的嘗試，其初步報告為首屆全球 AI 治理對話提供共同的證據基礎（Independent International Scientific Panel on AI, 2026），其專題簡報針對代理式 AI 的風險提出調整防護措施的建議（UN News, 2026b）。

企業提出安全主張，不代表其動機必然只有公共利益或商業考量。先進 AI 企業支持國際標準與外部評估，可能出於風險認知、聲譽管理、法律責任及參與規則制定等不同考量。本文所能檢驗的是公開發言、政策建議及政府回應，無法直接判定企業的內部動機。其承諾是否落實，仍應以獨立評估、資訊公開及後續行動作為依據。

然而，企業動機無法直接觀察，不表示其結構性影響可以忽略。管制經濟學指出，受管制產業可能藉由影響規則內容，使管制轉為保護既有業者，形成「管制俘虜」（regulatory capture）問題（Stigler, 1971; Dal Bó, 2006）。Wei 等人（2024）訪談 17 位 AI 政策專家後發現，受訪者最常提到的企業影響管道是議程設定與政策倡議，其次是學術影響與資訊管理。

從這個角度檢視先進 AI 企業所倡議的類似國際原子能總署機制、部署前測試及國際驗證，可見兩種可能性並存。一方面，若這類制度以高額合規成本為前提，資金與算力有限的新創企業及開源模型開發者將較難負擔，可能形成有利於既有大型企業的進入障礙。另方面，OpenAI 在 2023 年的倡議中主張，能

力門檻以下的企業與開源專案不應受到同等管制（Altman et al., 2023），Anthropic 在 2026 年也質疑美國政府的存取限制命令（Anthropic, 2026），顯示企業與政府的利益並非完全一致。兩種解讀的關鍵差異，在於門檻如何設定、由誰認定，以及合規成本由誰承擔。

因此，本文不將企業的安全論述直接視為管制俘虜，但認為門檻設定權本身就是值得關注的權力來源。巴基斯坦在安理會的發言（United Nations Security Council, 2026）顯示，這項疑慮已進入多邊外交場合。未來評估相關制度時，應檢視門檻是否會隨技術發展調整、評估者是否獨立於受評企業，以及中小型開發者與開發中國家能否實際參與規則制定。

五、聯合國體系中的 AI 安全治理

聯合國體系是觀察 AI 治理制度化最重要的場合。2023 年 7 月，安理會首度召開 AI 專題會議，標誌 AI 正式進入國際和平與安全的議程（United Nations Security Council, 2023）。接著，聯合國大會先後採納美國主導的第 78/265 號決議與中國主導的第 78/311 號決議（United Nations General Assembly, 2024a, 2024b），兩項決議雖均以共識採納，卻反映美中在聯合國框架內爭取 AI 議程設定權的競爭，前者強調安全、可靠且值得信賴的 AI，後者強調能力建設與縮小數位落差。

2024 年的高階諮詢機構報告，建議設立國際科學小組與全球政策對話等機制（United Nations High-level Advisory Body on Artificial Intelligence, 2024），此一建議在 2025 年經聯合國大會第 79/325 號決議付諸實現（United Nations General Assembly, 2025）。2026 年 7 月，科學小組發布初步報告，首屆全球 AI 治理對話亦於日內瓦舉行（UN News, 2026a），聯合國的 AI 治理軌道由此從構想走向實際運作。9 月的安理會高階簡報，使 AI 安全議題再度回到以國際和平與安全為職責的安理會（UN News, 2026b）。

然而，聯合國體系的 AI 治理仍有明顯限制。首先，科學小組與全球對話的功能在於提供證據與促進對話，並不具備制定拘束性規則的權限。另外，安理會的行動受制於常任理事國的否決權，在美中立場分歧的情況下，難以採納實質決議；2026 年 9 月的高階簡報即未產出任何成果文件，常任理事國之間對議題定位亦明顯分歧：法國主張比照原子能時代建立共同規範，英國強調政府不能將保護人民的責任外包給私人企業，美國表明國際對話不得走向全球治理，

俄羅斯質疑安理會處理該議題的適格性（United Nations Security Council, 2026）。再者，美國明確表態拒斥以全球架構管控 AI（Irwin, 2026），而中國在聯合國之外另行推動世界人工智慧合作組織，並已有 29 國簽署成立協定（Independent International Scientific Panel on AI, 2026）。

就目前的職權與政治條件而言，聯合國較能發揮議程協調及提供研究證據的作用，尚難成為直接執行全球 AI 管制的機構。各國另行參與不同的治理協議，也顯示未來可能維持多個安排並存的局面（Keohane & Victor, 2011; Cihon et al., 2020）。這些材料與命題三提出的跨境風險促成制度合作的預期相符，但制度能否有效降低風險，仍不能僅從機構成立推斷。

伍、人工智慧國際政治核心議題化的形成機制

一、技術能力與政治議題轉化

第三節與第四節的事件資料顯示，AI 的議題地位並非隨時間線性上升，而是在若干能力門檻被跨越時出現階段性躍升。本文將這個過程歸納為三個門檻。最初的門檻是「通用化」：當 AI 由特定任務工具轉變為可跨領域應用的通用能力時，其影響由產業政策延伸至國家整體競爭力，國家以國家戰略加以因應，2017 年中國的國家規劃即屬此一階段的代表（中華人民共和國國務院，2017）。接著的門檻是「規模化」：生成式 AI 在 2023 年大規模普及，使一般公眾直接感受到 AI 能力的躍進，風險論述隨之進入公共議程，並催生第一波多邊安全倡議。最後的門檻是「代理化」：AI 能夠自主規劃並執行多步驟行動，其風險由人為濫用延伸至系統越權，2025 年至 2026 年的網路間諜行動與評估事件即為例證（Anthropic, 2025; OpenAI, 2026）。

每跨越一道門檻，受影響的行為者範圍與潛在傷害類型都會擴大，政治化程度也隨之升高，此結果支持命題一。用議程設定理論的術語來說，能力門檻的跨越不斷為危機的累積與界定帶來新的指標與事件（Kingdon, 2011），使 AI 在國家決策者的注意力競爭中持續取得優先地位。

二、競爭機制：戰略化與安全化

競爭機制由戰略化與安全化兩個子過程構成。戰略化是指國家將 AI 及其生產要素界定為決定相對權力的戰略資產，並透過國家規劃、產業補貼與出口管制等工具加以掌控，其管制範圍由 2022 年的晶片與設備（Bureau of Industry and Security, 2022），擴展到 2026 年的先進模型（Anthropic, 2026）。安全化是指行為者將 AI 建構為生存威脅，並據以正當化超越常規的措施（Buzan et al., 1998）。

本文的一項重要觀察是，AI 議題呈現「雙軌安全化」的特徵。其一是「國家安全化」，指涉對象為國家的安全與相對權力，主要安全化行為者為國家政府，威脅來源被界定為競爭對手國的 AI 能力。美國將 AI 定位為必須贏得的競賽，並以國家安全為由限制晶片與模型的輸出（Bureau of Industry and Security, 2022; The White House, 2025），即屬此一軌道。其二是「人類安全化」，指涉對象為人類整體，主要安全化行為者為科學家與尖端 AI 企業負責人，威脅來源被界定為技術的失控與濫用。滅絕風險聲明、Yoshua Bengio 在安理會的示警，以及 Dario Amodei 關於 AI 可能對人類整體構成風險的陳述，均屬此一軌道（Center for AI Safety, 2023; Nichols et al., 2026; UN News, 2026b）。

兩軌安全化之間存在分歧。2026 年 9 月，美國在聯合國大會明確拒斥以全球架構管控 AI，並將 AI 定位為贏者全拿的競賽（Irwin, 2026），實際上是強化國家安全化、淡化人類安全化。美方代表在翌日的安理會會議主張，先進能力的快速進展並非暫停發展或以新的全球治理結構加以約束的理由，並呼籲各國拒絕「被恐懼統治」（United Nations Security Council, 2026），上述說法可視為針對人類安全化軌道的去安全化（desecuritization）論述（Wæver, 1995），尖端 AI 企業與科學家則選擇繞過本國政府，直接以安理會成員國為受眾，推動人類安全化。依 Balzacq（2005）的觀點，安全化的成效取決於受眾與脈絡，人類安全化軌道之所以在 2026 年獲得前所未有的迴響，關鍵正在於代理式 AI 越權事件提供具體可感的脈絡，使原本被視為科幻推演的威脅，轉化為可被政府官員親身確認的事實。上述發現支持命題二，並補充安全化理論對非國家安全化行為者的討論。

當兩條軌道在多邊場域交會時，何者更具主導力？根據 2025 年至 2026 年的資料，人類安全化軌道在議題設定與語彙提供方面較為成功，國家安全化軌道則掌握政策工具與資源分配。國家對非國家行為者所提出的安全主張，大致

採取三種回應方式。第一是吸納與改寫。英國於 2025 年 2 月將「AI 安全研究所」（AI Safety Institute）更名為「AI 安全保障研究所」（AI Security Institute），表明其將聚焦化學與生物武器、網路攻擊及犯罪濫用等具有安全意涵的風險，不再處理偏誤與言論自由問題（Department for Science, Innovation and Technology, 2025），美國商務部於同年 6 月將其 AI 安全研究所改組為「AI 標準與創新中心」（Center for AI Standards and Innovation, CAISI），保留對網路、生物與化學等「可證明風險」的評估，並強調防範外國政府對美國技術施加不必要的管制（U.S. Department of Commerce, 2025）。兩個案皆將原先以人類整體為指涉對象的安全問題，轉譯為國家可掌握與評估的安全問題，近似 Hansen（2012）所稱的重新表述。

第二是反制。美方在安理會呼籲拒絕「被恐懼統治」，巴基斯坦質疑「為先進定速」可能成為技術排除的工具（United Nations Security Council, 2026）。前者否定威脅的急迫性，後者質疑倡議者的動機，分別從大國與開發中國家的立場提出反安全化論述（Stritzel & Chang, 2015）。第三是選擇性採納。美中同意設置事故溝通管道（The White House, 2026b），採納科學界與企業長期倡議的事故通報概念，卻將其限縮於雙邊、政府對政府的形式，並未接受國際驗證機構等多邊主張。

整體而言，人類安全化訴求能否轉化為政策，取決於其能否被國家改寫為國家安全或可控的風險管理問題。就此而言，非國家行為者所推動的安全化，較接近 Corry（2012）所稱的風險化，其主要作用在於擴大預防性治理議程，而非直接正當化例外措施。至於企業推動此一軌道時，是否納入影響規則設計的考量，已於第四節第四小節討論。

三、合作機制：跨境風險與制度化

合作機制的起點是跨境風險產生的功能需求。第四節已指出，AI 風險在知識、管轄、執行與參與四個面向形成治理缺口，這些缺口無法由單一國家的國內法規填補，當風險的外溢性足夠明顯時，即便是彼此競爭的國家，也有誘因在特定領域建立協調機制。

就制度發展而言，本文觀察到三種路徑。最早出現的是「規範先行」路徑，由國際組織與倡議網絡提出原則性規範逐步擴散，落實為國際法的形式，OECD、UNESCO、歐盟與歐洲理事會的發展即屬此類（OECD, 2019; UNESCO,

2021; Council of Europe, 2024; Regulation (EU) 2024/1689, 2024）。另一種是「知識先行」路徑，由知識社群提供共同的風險評估，作為政府間協商的基礎，聯合國科學小組即為其制度化形式（Independent International Scientific Panel on AI, 2026）。還有一種是「窄域共識」路徑，即在大國競爭的情況下，選擇風險最高、利益最明確的特定領域達成有限協議，例如美中於 2026 年 9 月峰會同意建立的 AI 事件雙邊溝通管道（The White House, 2026b）。

三種路徑中，窄域共識路徑在大國競爭的格局下最具可行性，其原因在於，窄域協議不要求雙方放棄整體競爭優勢，只要求在災難性後果建立共同底線，此一邏輯與冷戰時期的軍備管制經驗相符，也與 Maas（2019）強調的可驗證管制標的相呼應。上述發現支持命題三。

四、競爭與合作的雙重動力

本文認為，競爭與合作未必彼此抵銷。大國競逐可能增加安全風險及合作需求，而新的合作安排也會引發規則主導權之爭。兩種作用共同影響 AI 的國際政治地位，表 4 比較其運作方式。

第一條迴路是「競爭引發合作需求」。大國之間的 AI 競賽加速先進能力的開發速度，而開發速度愈快，安全研究與防護措施愈難跟上，事故與越權的風險隨之升高，進而提高各方對合作機制的需求。Dario Amodei 主張放緩能力提升，其理由之一正是在商業誘因驅動的競爭下，風險可能被放大（Amodei, 2026）。Jervis（1978）描述的安全困境，在此轉化為一種「風險困境」：各方均擔心單方面放緩將使自身落後，卻又都承受競賽加速所帶來的共同風險。

第二條迴路是「合作場合成為競爭舞臺」。多邊制度的建立並未消弭競爭，反而使競爭轉移至規則制定權的爭奪。美中在聯合國大會分別主導不同取向的決議（United Nations General Assembly, 2024a, 2024b），中國在聯合國之外籌組世界人工智慧合作組織（董雪，2026），美國拒斥全球管控並強調本國領先地位（Irwin, 2026）。在此一迴路中，AI 治理成為大國展現領導力與爭取追隨者的工具。

表 4

競爭機制與合作機制之比較

比較面向	競爭機制	合作機制
核心邏輯	相對收益與權力平衡	共同風險與集體行動
子過程	戰略化、安全化（國家安全軌）	跨境風險、制度化
主要行為者	國家政府、國防與商務部門	國際組織、知識社群、部分尖端 AI 企業
代表性工具	國家規劃、出口管制、產業政策	宣言、決議、條約、科學小組、事故通報
指涉對象	國家安全與相對權力	人類整體安全與全球公共利益
對議題位階之影響	使 AI 進入高階政治	使 AI 成為多邊治理焦點
相互作用	競爭加速能力開發，擴大合作需求	合作場域轉為規則制定權之競逐

資料來源：作者自行整理。

這兩種作用顯示，大國即使持續競爭，也可能為避免嚴重事故而維持有限合作；但各方仍會爭取合作規則的制定權。本文以「競爭性合作」描述這種局面。它可以解釋命題四所預期的並行關係，卻不意味合作一定能緩和整體競爭。

五、AI 國際政治核心議題化模型

綜合前述分析，本文提出「AI 國際政治核心議題化模型」，將其發展過程分為五個相關環節。起初是「能力累積」，即 AI 跨越通用化、規模化與代理化三個門檻。接著是「議題論述框架」，即國家與非國家行為者分別以戰略競爭與人類安全兩種框架詮釋 AI。另外是「焦點事件觸發」，即具體的越權或濫用事件開啟政策之窗。隨後是「制度回應」，即國家與國際組織以管制措施、多邊決議與常設機制加以回應。最後是「位階固化」，即 AI 成為元首會談與安理會的常態議程，並與其他重大議題相互連結。五個環節並非單向推進，制度回應的不足與大國競爭的加劇，會回饋至能力累積與議題論述框架環節，形成持續強化的循環。

焦點事件在模型被視為可能的觸發條件，而非單獨的充分原因。2026 年的證據呈現三項相互銜接的現象：其一，OpenAI 與 Hugging Face 與澳洲事件使代理式 AI 越權風險具體化（Albanese, 2026; OpenAI, 2026）。其二，企業、科學界與多國領袖已提出獨立評估、事故通報與國際協調等政策方案。其三，相關事件與聯合國大會高階會期、安理會會議及美中元首會晤在時間上交會。這些時序與論述連結與 Kingdon（2011）的政策之窗機制相容，但仍需與會期巧合、現行制度累積及企業策略等競爭性解釋一併評估。

為使上述推論具備可檢驗性，以下依第一節預先界定的兩類證據逐項闡述。環形檢驗包含兩項條件。第一是時序。Anthropic 揭露的網路間諜行動於 2025 年 11 月公開（Anthropic, 2025），OpenAI 於 2026 年 8 月 26 日公布內部評估事件的調查結果（OpenAI, 2026），兩者均早於 9 月下旬的聯合國大會、安理會會議與美中峰會。澳洲事件於 9 月 24 日由總理證實（Albanese, 2026），亦早於白宮在 9 月 25 日公布設置事故溝通管道的事實清單（The White House, 2026b）。

第二是論述指涉。安理會簡報聚焦於失控 AI「真實且迫在眉睫」的威脅（UN News, 2026b）；芬蘭－挪威倡議要求共享並通報嚴重事故（Office of the President of the Republic of Finland, 2026）；美中則設置事故溝通管道。這些事實顯示，高層行為者確實指向同類風險。由於兩項條件均成立，焦點事件假設尚未被排除。然而，通過環形檢驗僅表示該假設仍可能成立，並不足以證明其成立（Mahoney, 2012）。

在確鑿證據檢驗方面，目前有兩項證據通過。第一，澳洲總理證實事件後，政府隨後啟動跨機關調查（Albanese, 2026）。決策者明確將政策行動連結至該事件，但由於調查對象正是該事件，這項證據雖具強證明力，適用範圍仍有限。第二，Dario Amodei 明確表示，遞迴式改進的加速與相關事件，是其主張放緩先進能力提升的理由（Amodei, 2026），因而構成企業層次的確鑿證據。至於國際層次，本文蒐集的美中官方文件與安理會紀錄中，尚未見決策者明言某一事件直接促成溝通管道的設置或會議的召開，因此這部分的確鑿證據檢驗尚未通過。

綜合而言，焦點事件在國內與非國家行為者層次的作用，獲得較強的證據支持，在國際層次則處於「通過環形檢驗、尚缺確鑿證據」的狀態。這說明本文為何將焦點事件定位為觸發條件，而非充分原因。

依據第二節所建立的判定規則，本文以 2026 年的經驗證據，逐項評估 AI 的議題地位（見表 5）。

表 5

AI 國際政治核心議題化之指標評估（2026 年）

指標	2026 年經驗證據	評估結果
議程層級	安理會高階簡報；美中元首會談納入 AI；Donald Trump 於聯合國大會一般性辯論專段論述 AI	高度
安全框架	國家以國家安全為由管制先進模型；科學家與企業負責人於安理會陳述人類安全威脅	高度
資源戰略化	管制範圍由晶片與設備擴及先進模型	高度
制度化密度	聯合國科學小組與全球對話正式運作；世界人工智慧合作組織成立協定簽署；歐洲理事會公約開放簽署；美中同意建立 AI 對話與事件溝通管道	高度
行為者擴散	國家、國際組織、尖端 AI 企業與科學界共同參與安理會議程	高度
跨議題連結	AI 與貿易、出口管制並列於美中峰會議程；與核武指揮管制相連結；美中八項成果將 AI 對話與關稅互減、禁毒合作並列	高度

資料來源：本表依表 3 之操作指標整理；各項證據出處詳見第一節至第四節。

依照本文設定的判定規則，2026 年 AI 在六項指標上均有符合門檻的證據，也達到「議程層級」和「安全框架」兩項必要條件，因此可列為國際政治的核心議題。這是依特定指標所做的分類，不代表其重要性已不可逆轉。判定是否穩健，仍有賴後續年度資料及其他國際議題的比較。

為檢驗上述判定反映的是持續性變化，而非判定規則過於寬鬆，本文以 2023 年作為基準對照。2023 年，AI 已進入安理會議程（United Nations Security Council, 2023），美中元首也於 Woodside Summit 同意建立 AI 政府間對話（Biden, 2023），議程層級可評為高度；滅絕風險聲明與 *The Bletchley Declaration* 使安全框架達到高度；安理會首場 AI 會議即邀請企業代表擔任簡報人，行為者擴散亦屬高度。

然而，當時的資源戰略化主要集中於晶片與製造設備，尚未擴及模型；制度化密度仍以宣言與原則為主，既無常設性的全球機制，也沒有具拘束力的國際條約；跨議題連結雖已出現，但 AI 仍只是峰會多項議程之一，約介於中度與高度。依表 3 的判定規則，AI 在 2023 年至多有四項指標達到高度門檻，應歸為「重要議題」而非「核心議題」。相較之下，2023 年至 2026 年間較明顯的變化，是資源戰略化延伸至最終產品，以及聯合國常設機制、歐洲理事會公約與美中 AI 對話等制度安排出現，這項結果與命題二、命題三的推論相符。

前述三項替代解釋均有一定說服力。聯合國大會會期、安理會輪值與元首互訪可以解釋為何相關活動集中於 2026 年 9 月，卻不足以說明事故通報與獨立評估等治理方案的出現。議題關注週期則提醒我們，單月的關注高峰不足以證明長期趨勢，因此本文另比較 2023 年的狀況，並納入 2024 年至 2026 年的制度發展。

企業可能藉安全論述影響政策規則，這項解釋與本文分析並不矛盾。安全化理論關心的是安全主張如何進入政治程序及其獲得何種回應，而不預設企業必然基於單一動機。現有資料顯示，焦點事件可能促成政策討論，但還不足以排除會期安排及既有治理倡議等其他因素。若未來一至二年新設機制沒有持續運作，AI 議題地位是否已趨穩定，也需要重新評估。

陸、結 論

本文以 AI 如何成為國際政治的核心議題為研究問題，結合安全化理論、權力政治與全球治理文獻，分析 2017 年至 2026 年的政策發展與主要事件。依研究架構及操作指標，得到以下發現。

一、AI 的通用目的性與代理自主能力，使它既可能增加國家的經濟與軍事實力，也可能產生跨境安全風險。前者引發戰略競爭，後者帶來單一國家難以獨力處理的治理問題。技術能力與應用範圍的擴張，為 AI 逐步進入國際政治議程提供了條件，但並不能單獨解釋政策變化。

二、競爭與合作可以並行，並相互影響。國家透過戰略規劃、安全論述及出口管制，使 AI 從產業政策進入高階外交；對跨境風險的擔憂，則促成原則倡議、國際協議與聯合國常設機制。競爭可能增加合作的必要，合作也可能成為爭奪國際規則主導權的途徑。本文以「競爭性合作」概括此種互動。

三、AI 安全論述大致沿著兩條途徑發展。國家政府主要關注相對實力與國家安全，科學家及先進 AI 企業則常以人類整體安全為訴求。2026 年，後者直接向安理會提出相關主張，顯示非國家行為者已參與高階國際安全議程；但安全主張是否獲政府接受，仍需依具體政策回應判斷。兩軌交會時，國家通常透過重新詮釋、反制或選擇性採納，回應非國家行為者的訴求。後者的影響更接近以「風險化」擴大預防性治理議程，而非直接促成例外措施。

四、2026 年代理式 AI 越權事件的公開，與高階外交討論在時間及風險論述上相近，符合焦點事件促成政策關注的部分預期。不過，事件與議程相繼發生並不能單獨證明因果關係；多年來的制度發展、會期安排及大國互動，也可能影響各方的政策選擇。

五、依六項操作指標，AI 於 2026 年符合本文所界定的「國際政治核心議題」。相較 2023 年，相關討論已擴及先進模型管制、常設國際機制與元首層級的事故溝通安排。然而，這項判斷僅適用本文設定的資料範圍與判定門檻，不能視為全球治理已達成穩定共識。

本文的貢獻就理論討論而言有三項：將權力政治與全球治理納入同一機制架構，說明競爭與合作何以能並行升高；以「雙軌安全化」區分以國家相對權力為指涉對象的安全化，以及以人類整體風險為指涉對象的安全化，並把尖端 AI 企業與科學家納入行為者分析；將國際政治核心議題轉為六項可觀察指標與判定規則。不過，這些貢獻屬於理論建構與單一議題的初步檢驗，其外部效度有待跨議題比較。

在經驗研究方面，本文整理 2017 年至 2026 年 9 月的重要政策與國際事件，補充既有文獻較少涵蓋的代理式 AI 越權案例、先進模型出口管制及聯合國新設治理機制。這些材料有助於說明議題如何變化，但近期事件仍有待後續調查與政策實踐確認。

就國際治理而言，本文的經驗分析支持「由可界定、可驗證的窄域問題著手，逐步累積合作」的方向，但尚不足以直接證成任何全面治理模式，建議可優先採取三項制度工具。首先，建立跨境 AI 事故通報機制，明確規範事件分類、通報時限、責任窗口與保密規則。美中已於 2026 年 9 月同意設置雙邊事故溝通管道，多國領袖亦要求共享嚴重安全事件資訊（Office of the President of the Republic of Finland, 2026; The White House, 2026b）。其次，發展共同測試基準、

第三方存取權及結果揭露規則，以提升先進模型評估的可驗證性。最後，針對 AI 輔助生化武器、核武指揮控制等高後果領域研議窄域協定。上述建議屬規範性推論，其可行性取決於大國是否願意接受互惠、查核與資訊保護安排。

本文對台灣而言有以下幾點政策意涵：先進半導體供應鏈地位使台灣受 AI 戰略化與供應鏈安全政策影響，宜持續強化算力治理、晶片可追溯性與出口管制合規能力；在無法參與聯合國機制的情況下，可透過學研網絡、產業標準組織與多邊合作，參與風險評估及技術標準制定；國內治理可優先建立政府資訊系統面對代理式 AI 越權存取時的偵測、通報與跨部會應變流程。澳洲事件顯示，政策問題不限於模型安全，也包括政府資訊系統能否及時辨識非典型自動化行為（Albanese, 2026）。

台灣的處境應置於「競爭性合作」的架構下加以檢視。根據波士頓顧問公司與美國半導體產業協會（Semiconductor Industry Association, SIA）於 2021 年發布的報告，全球 10 奈米以下先進邏輯晶片產能約有 92%位於台灣（Varas et al., 2021），從而台灣成為 AI 戰略資源化過程難以取代的關鍵節點。Farrell 與 Newman（2019）指出，掌握網絡樞紐的國家得以將相互依存關係轉化為影響力，惟其前提在於該國對樞紐具有管轄能力，且已建立運用該樞紐所需的國內制度。台灣所處條件則較為複雜，先進製程位於台灣，製程設備、設計軟體與相關智慧財產卻高度涉及美國技術，美國因此得以透過外國直接產品規則等域外管轄工具，介入台灣企業的出貨決策。

亦即，供應鏈的不可替代性更適宜作為累積互信、爭取參與規則討論的籌碼，而非可單獨操作的施壓手段，若其被外界認定為施壓工具，反可能加劇外界對台灣供應鏈風險的疑慮，促使各國加快產能分散。

美國出口管制對台灣產業造成的合規壓力，近年已具體顯現。2024 年 10 月，研究機構 TechInsights 拆解分析華為 Ascend 910B 晶片後，發現其中使用台積電的技術，台積電旋即向美方通報，美國商務部隨後要求台積電停止向中國客戶出貨 7 奈米及更先進製程的 AI 晶片，而華為據報係經由第三方間接取得相關晶片（Swain, 2024）。此事件顯示，即使台灣企業遵守本國法規，晶片仍可能透過第三方轉流至受管制實體，形成本文所稱的管轄缺口：產品在台灣製造，主要管制規範源自美國，違規流向則發生在第三地。

2025 年 1 月，美國公布《AI 擴散框架》，按目的地對各國實施分級管理，

以規範先進運算晶片與封閉模型權重的出口，台灣雖屬少數享有許可例外的夥伴，台灣廠商出口至其他地區的 AI 伺服器若含受管制的美國晶片，仍可能受該規則約束（Bureau of Industry and Security, 2025a）。該規則於同年 5 月廢止，美方並同步發布華為 Ascend 晶片使用風險與防範晶片轉運的指引（Bureau of Industry and Security, 2025b）。

對台灣企業而言，此情況意味著管制規則可能於短期內大幅變動，企業必須同時回應本國與美國的法規要求。台灣的因應途徑之一，在於使本國管制與主要夥伴接軌，經濟部國際貿易署於 2025 年 6 月將華為、中芯等 601 個實體列入戰略性高科技貨品出口實體管理名單，廠商須於出口前申請許可（中央社，2025）。

在無法參與聯合國機制的限制下，台灣較具可行性的路徑，是將技術能力轉化為可供其他國家採用的評估與標準工作。首先，立法院於 2025 年 12 月三讀通過《人工智慧基本法》，以國科會為主管機關，並要求數位發展部提供高風險 AI 的評估與驗證工具（蘇文彬，2025）。若此一國內評測能力得以在測試方法、事故分類與資料格式，與英國、美國及日本等國的 AI 安全相關機構對接，台灣可憑藉技術夥伴而非聯合國會員的身分，參與先進模型的評估網絡。

其次，晶片可追溯性是台灣具備比較優勢的領域。前述台積電案例顯示，掌握先進晶片的最終用途，是出口管制能否有效發揮功能的關鍵，台灣產業若能建立晶片識別、出貨稽核與客戶審查的共同作法，並透過國際半導體產業協會（Semiconductor Equipment and Materials International, SEMI）等產業組織推廣，即可將供應鏈地位轉化為參與規則制定的實際管道。

再者，政府資訊系統因應代理式 AI 越權存取風險時，可借鑑澳洲事件的經驗，建立跨部會的偵測與通報流程，並積極與理念相近國家交換事故資訊，此一作法並不以聯合國會員身分為前提，所面臨的政治阻力相對較低。

美中的「競爭性合作」亦可能使台灣承受雙重壓力。美中關係若進一步緊張，台灣可能須被要求在出口管制與供應鏈布局配合更嚴格的規定，美中若就 AI 風險議題形成窄域協議，台灣則可能在未被諮詢的情況下承擔該協議的外溢效應，例如提高晶片流向的監控要求。對此，台灣宜採行更具彈性的敏捷治理模式，譬如善用《人工智慧基本法》納入的創新實驗環境（蘇文彬，2025），藉由漸進式法規調適框架及定期檢討機制，回應管制規則的快速變動，在出口

管制確保與主要夥伴維持相容，並透過學研網絡與產業標準組織持續投入技術討論。

前述建議係屬政策推論，可行性仍有賴於美中關係的走向與國內法制的配合程度。

本文存在四項研究限制。首先，觀察期間仍然有限。2026 年的多項事件尚處於調查或制度建構階段，美中事故溝通管道的實際運作情形、澳洲調查結論，以及新設多邊機制能否延續，現階段均未確定。其次，資料來源的完整程度不同。企業自我揭露可能存在選擇性，而國家安全決策經常受到機密限制。另外，案例主要聚焦於美中與聯合國體系，對歐盟、印度、中東及全球南方的比較分析仍顯不足。最後，六項指標及「至少五項且包含議程層級與安全框架」之門檻，皆為本文提出的測量規則，尚未經跨議題驗證。因此，本文對核心議題的判定，應視為可供檢驗的研究命題，而非無條件成立的客觀分類。

為回應上述限制，未來研究可優先處理三項工作，包括針對聯合國大會、安理會與主要國家外交文件進行有系統的文本分析，檢驗 AI 議程重要性在 2026 年後是否持續；將六項指標套用於核武、氣候變遷與網路安全等議題，以校準門檻並檢驗外部效度；追蹤聯合國科學小組、全球 AI 治理對話、世界人工智慧合作組織及美中事故溝通機制的實際運作，以區分「制度成立」與「制度有效」兩種結果。

參考文獻

中央社（2025）。經濟部出手！華為、中芯等 601 實體涉「武擴」遭列出口黑名單：台商要出貨得申請許可。**數位時代**，6 月 16 日。https://www.bnext.com.tw/article/83578/huawei-smic-forbidden [Central News Agency (2025). Jingjibu chushou! Huawei, Zhongxin deng 601 shiti she “wukuo” zao lie chukou heimingdan: Taishang yao chuhuo dei shenqing xuke. *Business Next*, June 16.]

中華人民共和國外交部（2021）。**中國關於規範人工智能軍事應用的立場文件**，12 月 14 日。https://www.mfa.gov.cn/web/ziliao_674904/zcwj_674915/202112/t20211214_10469511.shtml [Ministry of Foreign Affairs of the People’s Republic of China (2021). *Position paper of the People’s Republic of China on regulating military applications of artificial intelligence (AI)*, December 14.]

中華人民共和國外交部（2023）。**全球人工智能治理倡議**，10 月 20 日。https://www.mfa.gov.cn/web/ziliao_674904/1179_674909/202310/t20231020_11164831.shtml [Ministry of Foreign Affairs of the People's Republic of China (2023). *Global AI Governance Initiative*, October 20.]

中華人民共和國外交部（2026）。**習近平同美國總統特朗普會談**，9 月 25 日。https://www.mfa.gov.cn/zyxw/202609/t20260925_12031134.shtml [Ministry of Foreign Affairs of the People's Republic of China (2026). *Xi Jinping tong Meiguo zongtong Telangpu huitan*, September 25.]

中華人民共和國國務院（2017）。**國務院關於印發新一代人工智能發展規劃的通知**（國發〔2017〕35 號）。中華人民共和國中央人民政府，7 月 20 日。https://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm [State Council of the People's Republic of China (2017). *Guowuyuan guanyu yinfa xin yidai rengong zhineng fazhan guihua de tongzhi* (Guofa 2017 No. 35). Central People's Government of the People's Republic of China, July 20.]

世界人工智能大會暨人工智能全球治理高級別會議（2025）。**人工智能全球治理行動計劃**。中華人民共和國外交部，7 月 26 日。https://www.fmprc.gov.cn/web/wjb_673085/zzjg_673183/xws_674681/xgxw_674683/202507/t20250726_11677803.shtml [World Artificial Intelligence Conference and High-Level Meeting on Global AI Governance (2025). *Global AI Governance Action Plan*. Ministry of Foreign Affairs of the People's Republic of China, July 26.]

米勒（Miller, C.）（2023）。**晶片戰爭：矽時代的新賽局，解析地緣政治下全球最關鍵科技的創新、商業模式與台灣的未來**（洪慧芳譯）。天下雜誌。（原著出版於 2022）[Miller, C. (2023). *Chip war: The fight for the world's most critical technology* (Hung, Hui-Fang, Trans.). CommonWealth Magazine. (Original work published 2022)]

封帥（2018）。人工智能時代的國際關係：走向變革且不平等的世界。**外交評論（外交學院學報），35**（1），128-156。[Feng, Shuai (2018). Rengong zhineng shidai de guoji guanxi: Zouxiang biange qie bu pingdeng de shijie. *Foreign Affairs Review, 35*(1), 128-156.]

傅瑩（2019）。人工智能對國際關係的影響初析。**國際政治科學，4**（1），1-18。[Fu, Ying (2019). Rengong zhineng dui guoji guanxi de yingxiang chuxi. *Quarterly Journal of International Politics, 4*(1), 1-18.]

董雪（2026）。**外交部：將推動世界人工智能合作組織儘早啟動運轉**。新華網，7 月 17 日。https://www.news.cn/world/20260717/d8109ef663604ef68ba99bf9085d5afd/c.html [Dong, Xue (2026). *Waijiaobu: Jiang tuidong shijie rengong zhineng hezuo zuzhi jinzao qidong yunzhuan*. Xinhuanet, July 17.]

蘇文彬（2025）。**立院三讀通過 AI 基本法，明定國科會為基本法主管機關、行政院成立國家 AI 戰略特別委員會**。iThome，12 月 23 日。https://www.ithome.com.tw/news/172980 [Su, Wen-Bin (2025). *Liyuan sandu tongguo AI jibenfa, mingding Guokehui wei jibenfa zhuguan jiguan, Xingzhengyuan chengli guojia AI zhanlüe tebie weiyuanhui*. iThome, December 23.]

Abbott, K. W., & Snidal, D. (2000). Hard and soft law in international governance. *International Organization, 54*(3), 421-456. https://doi.org/10.1162/002081800551280

Adcock, R., & Collier, D. (2001). Measurement validity: A shared standard for qualitative and quantitative research. *American Political Science Review, 95*(3), 529-546. https://doi.org/10.1017/S0003055401003100

Albanese, A. (2026). *Press conference – New York* [Transcript]. Prime Minister of Australia, September 24. https://www.pm.gov.au/media/press-conference-new-york

Allison, G. (2017). *Destined for war: Can America and China escape Thucydides's trap?* Houghton Mifflin Harcourt.

Altman, S., Brockman, G., & Sutskever, I. (2023). *Governance of superintelligence*. OpenAI, May 22. https://openai.com/index/governance-of-superintelligence/

Amodei, D. (2026). *We must pace the frontier*, September 12. https://darioamodei.com/post/we-must-pace-the-frontier

Anthropic. (2025). *Disrupting the first reported AI-orchestrated cyber espionage campaign*, November 13. https://www.anthropic.com/news/disrupting-AI-espionage

Anthropic. (2026). *Statement on the US government directive to suspend access to Fable 5 and Mythos 5*, June 12. https://www.anthropic.com/news/fable-mythos-access

Balzacq, T. (2005). The three faces of securitization: Political agency, audience and context. *European Journal of International Relations, 11*(2), 171-201.

https://doi.org/10.1177/1354066105052960

Bengio, Y., Hinton, G., Yao, A., Song, D., Abbeel, P., Darrell, T., Harari, Y. N., Zhang, Y.-Q., Xue, L., Shalev-Shwartz, S., Hadfield, G., Clune, J., Maharaj, T., Hutter, F., Baydin, A. G., McIlraith, S., Gao, Q., Acharya, A., Krueger, D., . . . Mindermann, S. (2024). Managing extreme AI risks amid rapid progress. *Science, 384*(6698), 842-845. https://doi.org/10.1126/science.adn0117

Biden, J. R. (2023). *Readout of President Joe Biden's meeting with President Xi Jinping of the People's Republic of China*. The American Presidency Project, November 15. https://www.presidency.ucsb.edu/node/367862

The Bletchley Declaration by countries attending the AI Safety Summit, 1–2 November 2023. (2023). GOV.UK, November 1. https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023

Bradford, A. (2023). *Digital empires: The global battle to regulate technology*. Oxford University Press.

Bresnahan, T. F., & Trajtenberg, M. (1995). General purpose technologies 'Engines of growth'? *Journal of Econometrics, 65*(1), 83-108. https://doi.org/10.1016/0304-4076(94)01598-T

Bureau of Industry and Security. (2022). *Commerce implements new export controls on advanced computing and semiconductor manufacturing items to the People's Republic of China (PRC)* [Press release]. U.S. Department of Commerce, October 7. https://www.bis.gov/press-release/commerce-implements-new-export-controls-advanced-computing-semiconductor-manufacturing-items-peoples

Bureau of Industry and Security. (2025a). Framework for artificial intelligence diffusion. *Federal Register, 90*(9), 4544-4584. https://www.govinfo.gov/content/pkg/FR-2025-01-15/pdf/2025-00636.pdf

Bureau of Industry and Security. (2025b). *Department of Commerce announces rescission of Biden-era artificial intelligence diffusion rule, strengthens chip-related export controls* [Press release]. U.S. Department of Commerce, May 13. https://www.bis.gov/press-release/department-commerce-announces-rescission-biden-era-artificial-intelligence-diffusion-rule-strengthens

Buzan, B., Wæver, O., & de Wilde, J. (1998). *Security: A new framework for analysis*. Lynne Rienner.

Center for AI Safety. (2023). *Top AI scientists warn: Risk of extinction from AI on scale with nuclear war* [Press release], May 30. https://www.safe.ai/work/press-release-ai-risk

Cihon, P., Maas, M. M., & Kemp, L. (2020). Fragmentation and the future: Investigating architectures for international AI governance. *Global Policy, 11*(5), 545-556. https://doi.org/10.1111/1758-5899.12890

Collier, D. (2011). Understanding process tracing. *PS: Political Science & Politics, 44*(4), 823-830. https://doi.org/10.1017/S1049096511001429

Corry, O. (2012). Securitisation and 'riskification': Second-order security and the politics of climate change. *Millennium: Journal of International Studies, 40*(2), 235-258. https://doi.org/10.1177/0305829811419444

Council of Europe. (2024). *Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law* (Council of Europe Treaty Series No. 225).

Crouse, M. (2025). *US and UK decline 'inclusive and sustainable' Paris AI summit declaration*. eWeek, February 11. https://www.eweek.com/?p=232240

Dal Bó, E. (2006). Regulatory capture: A review. *Oxford Review of Economic Policy, 22*(2), 203-225. https://doi.org/10.1093/oxrep/grj013

Department for Science, Innovation and Technology. (2025). *Tackling AI security risks to unleash growth and deliver Plan for Change* [Press release]. GOV.UK, February 14. https://www.gov.uk/government/news/tackling-ai-security-risks-to-unleash-growth-and-deliver-plan-for-change

Ding, J. (2024). *Technology and the rise of great powers: How diffusion shapes economic competition*. Princeton University Press.

Ding, J., & Dafoe, A. (2023). Engines of power: Electricity, AI, and general-purpose, military transformations. *European Journal of International Security, 8*(3), 377-394. https://doi.org/10.1017/eis.2023.1

Downs, A. (1972). Up and down with ecology: The "issue-attention cycle." *The Public Interest, 28*, 38-50.

Drezner, D. W. (2019). Technological change and international relations. *International Relations, 33*(2), 286-303. https://doi.org/10.1177/0047117819834629

Durkee, A. (2026). Bill Gates warns AI is ‘powerful’ enough to ‘cause a billion deaths’. *Forbes*, September 25. https://www.forbes.com/sites/alisondurkee/2026/09/25/bill-gates-warns-ai-is-powerful-enough-to-cause-a-billion-deaths/

Exec. Order No. 14110, 88 Fed. Reg. 75191 (2023).

Exec. Order No. 14179, 90 Fed. Reg. 8741 (2025).

Farrell, H., & Newman, A. L. (2019). Weaponized interdependence: How global economic networks shape state coercion. *International Security, 44*(1), 42-79. https://doi.org/10.1162/isec_a_00351

Finnemore, M., & Sikkink, K. (1998). International norm dynamics and political change. *International Organization, 52*(4), 887-917. https://doi.org/10.1162/002081898550789

Future of Life Institute. (2023). *Pause giant AI experiments: An open letter*, March 22. https://futureoflife.org/open-letter/pause-giant-ai-experiments/

George, A. L., & Bennett, A. (2005). *Case studies and theory development in the social sciences*. MIT Press.

Goertz, G. (2006). *Social science concepts: A user’s guide*. Princeton University Press.

Haas, E. B. (1980). Why collaborate? Issue-linkage and international regimes. *World Politics, 32*(3), 357-405. https://doi.org/10.2307/2010109

Haas, P. M. (1992). Introduction: Epistemic communities and international policy coordination. *International Organization, 46*(1), 1-35. https://doi.org/10.1017/S0020818300001442

Handley, E. (2026). *OpenAI hacked Medicare portal, Prime Minister Anthony Albanese says*. ABC News, September 24. https://www.abc.net.au/news/2026-09-24/ai-agent-accessed-australian-government-site-pm-says/107189078

Hansen, L. (2012). Reconstructing desecuritisation: The normative-political in the Copenhagen School and directions for how to apply it. *Review of International Studies, 38*(3), 525-546. https://doi.org/10.1017/S0260210511000581

Hansen, L., & Nissenbaum, H. (2009). Digital disaster, cyber security, and the Copenhagen School. *International Studies Quarterly, 53*(4), 1155-1175. https://doi.org/10.1111/j.1468-2478.2009.00572.x

Hoffmann, S. (1966). Obstinate or obsolete? The fate of the nation-state and the case of Western Europe. *Daedalus, 95*(3), 862-915.

Horowitz, M. C. (2018). Artificial intelligence, international competition, and the balance of power. *Texas National Security Review, 1*(3), 36-57. https://doi.org/10.15781/T2639KP49

Independent International Scientific Panel on AI. (2026). *Preliminary report of the Independent International Scientific Panel on AI: Evidence-based assessment of opportunities, risks and impacts of AI*. United Nations. https://www.un.org/independent-international-scientific-panel-ai/en/preliminary-report

Irwin, L. (2026). Trump talks Iran war, new 'super intelligence' in United Nations speech. *Deseret News*, September 22. https://www.deseret.com/politics/2026/09/22/president-trump-delivers-remarks-at-united-nations-general-assembly-meeting-about-iran-war-and-artificial-intelligence/

Jervis, R. (1978). Cooperation under the security dilemma. *World Politics, 30*(2), 167-214. https://doi.org/10.2307/2009958

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*(9), 389-399. https://doi.org/10.1038/s42256-019-0088-2

Keck, M. E., & Sikkink, K. (1998). *Activists beyond borders: Advocacy networks in international politics*. Cornell University Press.

Keohane, R. O. (1984). *After hegemony: Cooperation and discord in the world political economy*. Princeton University Press.

Keohane, R. O., & Nye, J. S. (1977). *Power and interdependence: World politics in transition*. Little, Brown.

Keohane, R. O., & Victor, D. G. (2011). The regime complex for climate change. *Perspectives on Politics, 9*(1), 7-23. https://doi.org/10.1017/S1537592710004068

Kingdon, J. W. (2011). *Agendas, alternatives, and public policies* (Updated 2nd ed.). Longman.

Maas, M. M. (2019). How viable is international arms control for military artificial intelligence? Three lessons from nuclear weapons. *Contemporary Security Policy, 40*(3), 285-311. https://doi.org/10.1080/13523260.2019.1576464

Mahoney, J. (2012). The logic of process tracing tests in the social sciences. *Sociological Methods & Research, 41*(4), 570-597.

https://doi.org/10.1177/0049124112437709

Ministry of Foreign Affairs of the People's Republic of China. (2026). *China and the United States reach eight deliverables and understandings*, September 26. https://www.fmprc.gov.cn/eng/xw/zyxw/202609/t20260926_12031663.html

Nichols, M., Irish, J., & Rozen, C. (2026). *AI leaders warn UN of security risks as systems grow more powerful*. CP24, September 23. https://www.cp24.com/news/world/2026/09/23/ai-leaders-warn-un-of-security-risks-as-systems-grow-more-powerful/

OECD. (2019). *Recommendation of the Council on Artificial Intelligence* (OECD/LEGAL/0449). https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

Office of the President of the Republic of Finland. (2026). *A call for control of frontier AI models*, September 21. https://www.presidentti.fi/en/a-call-for-control-of-frontier-ai-models/

OpenAI. (2026). *The Hugging Face incident and the road ahead*, August 26. https://openai.com/index/hugging-face-incident-and-the-road-ahead/

Payne, K. (2018). Artificial intelligence: A revolution in strategic affairs? *Survival, 60*(5), 7-32. https://doi.org/10.1080/00396338.2018.1518374

Power, J. (2026). *US lifts restrictions on Anthropic's powerful AI models Fable and Mythos*. Al Jazeera, July 1. https://www.aljazeera.com/economy/2026/7/1/us-lifts-restrictions-on-powerful-ai-models-fable-mythos-anthropic-says

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). (2024). *Official Journal of the European Union*, L, 2024/1689. http://data.europa.eu/eli/reg/2024/1689/oj

Roberts, H., Hine, E., Taddeo, M., & Floridi, L. (2024). Global AI governance: Barriers and pathways forward. *International Affairs, 100*(3), 1275-1286. https://doi.org/10.1093/ia/iiae073

Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Viking.

Scharre, P. (2023). *Four battlegrounds: Power in the age of artificial intelligence*. W.

W. Norton & Company.

Schmitt, L. (2022). Mapping global AI governance: A nascent regime in a fragmented landscape. *AI and Ethics, 2*(2), 303-314. https://doi.org/10.1007/s43681-021-00083-y

Stigler, G. J. (1971). The theory of economic regulation. *The Bell Journal of Economics and Management Science, 2*(1), 3-21. https://doi.org/10.2307/3003160

Stritzel, H., & Chang, S. C. (2015). Securitization and counter-securitization in Afghanistan. *Security Dialogue, 46*(6), 548-567. https://doi.org/10.1177/0967010615588725

Swain, G. (2024). *US orders TSMC to halt advanced chip exports to China*. Computerworld, November 11. https://www.computerworld.com/article/3602389/us-orders-tsmc-to-halt-advanced-chip-exports-to-china.html

Taddeo, M., & Floridi, L. (2018). Regulate artificial intelligence to avert cyber arms race. *Nature, 556*(7701), 296-298. https://doi.org/10.1038/d41586-018-04602-6

Tallberg, J., Erman, E., Furendal, M., Geith, J., Klamberg, M., & Lundgren, M. (2023). The global governance of artificial intelligence: Next steps for empirical and normative research. *International Studies Review, 25*(3), Article viad040. https://doi.org/10.1093/isr/viad040

UN News. (2026a). *Global push for AI governance amid warnings of 'catastrophic harm'*. United Nations, July 5. https://news.un.org/en/story/2026/07/1167862

UN News. (2026b). *LIVE: OpenAI and Anthropic brief Security Council amid 'real and imminent' threat posed by runaway AI*. United Nations, September 23. https://news.un.org/en/story/2026/09/1168414

UNESCO. (2021). *Recommendation on the ethics of artificial intelligence*.

United Nations General Assembly. (2024a). *Enhancing international cooperation on capacity-building of artificial intelligence* (A/RES/78/311). https://undocs.org/A/RES/78/311

United Nations General Assembly. (2024b). *Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development* (A/RES/78/265). https://undocs.org/A/RES/78/265

United Nations General Assembly. (2025). *Terms of reference and modalities for the*

establishment and functioning of the Independent International Scientific Panel on Artificial Intelligence and the Global Dialogue on Artificial Intelligence Governance (A/RES/79/325). https://undocs.org/A/RES/79/325

United Nations High-level Advisory Body on Artificial Intelligence. (2024). *Governing AI for humanity: Final report*. United Nations. https://www.un.org/sites/un2.un.org/files/governing_ai_for_humanity_final_report_en.pdf

United Nations Security Council. (2023). *Maintenance of international peace and security: Artificial intelligence: Opportunities and risks for international peace and security* (S/PV.9381). United Nations, July 18. https://undocs.org/S/PV.9381

United Nations Security Council. (2026). *Artificial intelligence and international security* (10228th meeting) [Automated transcript]. United Nations, September 23. https://transcripts.un.org/en/sc/10228

Urbina, F., Lentzos, F., Invernizzi, C., & Ekins, S. (2022). Dual use of artificial-intelligence-powered drug discovery. *Nature Machine Intelligence, 4*(3), 189-191. https://doi.org/10.1038/s42256-022-00465-9

U.S. Department of Commerce. (2025). *Statement from U.S. Secretary of Commerce Howard Lutnick on transforming the U.S. AI Safety Institute into the pro-innovation, pro-science U.S. Center for AI Standards and Innovation*, June 3. https://www.commerce.gov/node/7172

Varas, A., Varadarajan, R., Goodrich, J., & Yinug, F. (2021). *Strengthening the global semiconductor supply chain in an uncertain era*. Boston Consulting Group & Semiconductor Industry Association. https://www.semiconductors.org/wp-content/uploads/2021/05/BCG-x-SIA-Strengthening-the-Global-Semiconductor-Value-Chain-April-2021_1.pdf

Wæver, O. (1995). Securitization and desecuritization. In R. D. Lipschutz (Ed.), *On security* (pp. 46-86). Columbia University Press.

Wei, K., Ezell, C., Gabrieli, N., & Deshpande, C. (2024). How do AI companies "fine-tune" policy? Examining regulatory capture in AI governance. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 7*(1), 1539-1555. https://doi.org/10.1609/aies.v7i1.31745

The White House. (2025). *Winning the race: America's AI action plan*. https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-

Action-Plan.pdf

The White House. (2026a). *President Trump at the United Nations: "While others have talked, I have acted,"* September 22. https://www.whitehouse.gov/releases/2026/09/president-trump-at-the-united-nations-while-others-have-talked-i-have-acted/

The White House. (2026b). *Fact sheet: President Donald J. Trump advances a fair and reciprocal relationship with China while hosting historic state visit*, September 25. https://www.whitehouse.gov/fact-sheets/2026/09/fact-sheet-president-donald-j-trump-advances-a-fair-and-reciprocal-relationship-with-china-while-hosting-historic-state-visit/

Why Has Artificial Intelligence Become a Central Issue in International Politics?

William Yang*

Abstract

Artificial intelligence (AI) has moved beyond industrial and technology policy to become an important subject of summit diplomacy, the United Nations Security Council, and international governance. This article draws on securitization theory, power politics, and global governance scholarship to examine how technological capabilities, competition, cooperation, and specific incidents have affected AI's political standing. It assesses that standing using six indicators: diplomatic agenda level, security framing, strategic control of resources, institutional development, the range of participating actors, and links to other policy issues. The analysis combines documentary research with process tracing of official records, intergovernmental documents, corporate disclosures, and news reports published between 2017 and September 2026. The findings indicate that AI's general-purpose applications and growing capacity for autonomous action have made it both a resource in interstate competition and a potential source of risks extending across national borders. Strategic planning, export controls, and national security arguments have increased its diplomatic importance. Concerns about cross-border incidents have also encouraged governments, international organizations, scientists, and technology companies to pursue common safeguards. Their claims follow two broadly distinguishable paths, emphasizing national security and human security respectively. Incidents involving AI agents in 2026 coincided with prominent diplomatic discussions and shared concerns about inadequate oversight. This evidence is consistent with, but does not establish, a causal role for focusing events. Under the article's explicit criteria, AI qualifies as a central issue in international politics in 2026. This judgment remains provisional because investigations are ongoing and the thresholds have yet to be tested against other policy issues. The article discusses incident notification, independent safety assessments, limited international agreements, and policy implications for Taiwan.

Keywords: Artificial Intelligence, International Politics, Securitization, Global Governance, US–China Technological Competition

* William Yang received his Ph.D. in Political Science and International Relations from University of Warwick (U.K.) and is currently Professor in the Department of Public Administration and Management at Shih Hsin University. His research expertise includes AI in Social Sciences, Trumpism, Global Governance, International Climate Politics. Email: william@mail.shu.edu.tw.
The paper was published under two double-blind reviews. Received: October 1, 2026. Accepted: October 26, 2026.