參、 領域專題報告：

「以 Claude Constitution 發展為中心-以多方利害關係概念建立具備永續性的人工智慧治理模式」

一、 摘要

建構人工智慧治理模式，與單純訂定人工智慧新法並不相同[1]。人工智慧治理本身存在多元的治理模式，涵蓋標準規範、 產業自律治理、契約治理與規範、倫理委員會與治理機構、國際協議與制度、技術專業治理等；而在上位概念的治理模式中，也就包含訂定專法，以作為一種全新、獨立約束力的治理模式，這也就是各國制定人工智慧專法或基本法的原因[2]。

然而如溯及至堪稱早期、完整制定的人工智慧專法，要屬於歐盟在 2024 年正式通過的人工智慧法（ “EU AI Act”），其以風險分級導向的監管措施及制度為主，並依人工智慧系統涉及之風險程度配置不同規範要求。此一模式代表人工智慧治理由早期以倫理原則、自律規範及既有法律適用逐漸轉向專法規範。然而，EU AI Act 的出現並未減緩「AI 是否需要全新法律規範」相關爭議，更使問題進一步轉化，顯現出涵蓋人工智慧專法之規範內容、分類風險本身的可行性，以及法律如何因應技術的快速變化[3]等議題。亦即在人工智慧科技持續發展及時間考驗的經過後，若謹守以風險分類或是事後諸葛的角度來觀察各類新型態、具備自主決策力、規劃與執行能力的「代理型人工智慧」，將可能使法律規範之定位成為無法趕上科技變化的腳步[4]。亦即，針對這些已

---

1 Veale, M., Matus, K., & Gorwa, R. (2023). *AI and Global Governance: Modalities, Rationales, Tensions*, ANNU. REV. LAW SOC. SCI. 19(1), 255–275. https://doi.org/10.1146/annurev-lawsocsci-020223-040749 (Last visited: Sep 20, 2026).

2 Gstrein, Oskar J., Haleem, N. & Zwitter, A. (2024). *General-purpose AI regulation and the European Union AI Act*, INTERNET POLICY REV. 13(3). https://policyreview.info/articles/analysis/general-purpose-ai-regulation-and-ai-act (Last visited: Sep 20, 2026).

3 Ballot Jones, L., Thornton, J. & De Silva, D.(2025). *Limitations of risk-based artificial intelligence regulation: a structuration theory approach.* DISCOV ARTIF INTELL 5, 14. https://doi.org/10.1007/s44163-025-00233-9 (Last visited: Sep 20, 2026).

4 Ibid.

經佔據人類生活相當高比例的人工智慧科技，作為未來發展所不可或缺的得力助手，顯見人工智慧似乎作為一種「獨立個體」而存在，不論何種領域、或是各類型思考決斷的過程，其都存在相當廣泛且直接的影響力。而且科技本身亦可能作為地緣政治下的重要競爭資源[5]。在在顯示著專法本身似乎並非全面應對人工智慧相關規範的有效選擇。

本文因此嘗試藉由其他法學概念引導多邊利害關係治理概念，提出新型態的規範模式，認為人工智慧相關規範並不局限於「專法」或「特定領域」法律之探討，而是更明確的說，與其控制人工智慧所造成的風險，人工智慧的核心風險涉及的是一個獨立命題，也就是建構人工智慧其本身運作依據的具體模型中仰賴的原理原則授權一句從何而來？在 2026 年 1 月，Anthropic 以創新概念宣布將建制大型語言模型的直接憲法原則，也就是 Claude Constitution[6]。與傳統上依賴大量「人工標註的 RLHF（人類反饋強化學習）」不同，人工智慧憲法原則的基本概念是，透過提供整組明確「憲法原則」（例如：有效性、誠實信用、無害等），使 AI 在產出過程中，不斷進行自我批評與修正。而這類命題的關鍵就在於，人工智慧世界中的原理原則，究由誰予授權更為適合？[7]

首先，針對私部門本身，就有許多不同治理主體的規範想像。本文主要以科技產業龍頭的論點為主進行質性內容分析，包含黃仁勳所代表的 NVIDIA 提出了產業自律的觀點，在 2026 年 9 月主張，人工智慧的安全或規範，是工程系統與設計本身的課題，企業將對自己的產品負責，因此其認為並沒有必要再創設新興的 AI 法規[8]。而近期 OpenAI 在政策文件方面則表示支持對 frontier AI 建立登錄、揭露、甚至發給證照等等機制，並進而主張

---

5 David U. Socol de la Osa (Oct 13, 2025). *The Geopolitics of AI Governance: From Cooperation to Competition, and a Path to Stability*, NUS Centre for Research in Law & Technology. https://blog.nus.edu.sg/lawresearch/2025/10/13/the-geopolitics-of-ai-governance-from-cooperation-to-competition-and-a-path-to-stability/ (Last visited: Sep 20, 2026).

6 Frazier, K. (Jan 21, 2026). *Interpreting Claude's constitution*. Lawfare, https://www.lawfaremedia.org/article/interpreting-claude-s-constitution (Last visited: Sep 20, 2026).

7 Gstrein, Haleem, & Zwitter. Supra note 2.

8 Bort, J. (Sep 16, 2026). *We don't need AI regulation — leave safety to us, Nvidia's Jensen Huang says*, TechCrunch. https://techcrunch.com/2026/09/15/we-dont-need-ai-regulation-leave-safety-to-us-nvidias-jensen-huang-says/?trk=article-ssr-frontend-pulse_little-text-block&utm (Last visited: Sep 20, 2026).

政府、產業與其他利害關係人應可共同制定標準。但結論而言，私部門角色即使並未統一立場，共同特仍是色作為最直接的技術專業、發展潛力及實際控制能力，此等開發者所傳達出的理念更接近於內部倫理委員會或治理機制、示警系統、技術標準與模型評估等等直接規範架構模式，也呼應了人工智慧憲法的建構可能性。Microsoft、IBM、Google 等企業更均曾建立 ethics boards、responsible offices 或相關倫理治理機構。這顯示企業不只是受法律規範的對象，也逐漸成為實際產生人工智慧規範的制定者（norm producers）。

而政府所主張的治理邏輯通常建立在民主授權、法律拘束力、公共利益及基本權保障之上。EU AI Act 就是目前最具代表性的例子。歐盟認為特定技術特徵可能產生既有法律難以充分處理的健康、安全及基本權風險，因此建立一套全新風險分級為核心的法律框架。雖然如前所述，專法的治理架構本身並非完足，尤值關注的論點在於，以歐盟治理模式為例所提出的並不是將法律作為最後手段，而是讓法律建立共同規範底線，再透過技術標準、Code of Practice、行政機關與市場監督等機制具體落實。

最後，回到受有直接影響及間接影響的人工智慧服務使用者之角色分析；在公部門具有正當授權基礎的立法者往往無法掌握核心關鍵技術，然而私部門所擁有的專業卻掌控於特定群體，因而最終，更有賴私法中 Fiduciary Duty 的概念作為建構基礎，串連授權的意義與技術專業參與的可能性，成為公民社會與一般使用者作為受影響最深的關係人真正參與建置規範的契機。

二、 研究目的

既有的人工智慧專法，仍然存在許多無法因應科技快速變化的缺口，顯示以專法或補充規範的治理模式存在不足。人工智慧的運作與價值判斷本身，深刻影響著使用者與社會，因此這些價值由誰決定、是否具備正當性、透過何種治理機制或程序則成為關鍵問題。

作為引領未來科技發展的潮流之一，人工智慧的相關法律規範議題，往往大部分圍繞在是否具有新法制定的必要？然而隨著科技發展速度往往大幅超越法律制定速度的現象普遍存在，將切入角度轉換為建構技術本身為出發，脫離針對風險進行應對的思考模式，成為可發展的規範焦點。

人工智慧的相關議題看似涵蓋面向多元，實際、直接與間接之影響更難以窮盡，但若仔細觀察，能夠發現人工智慧的法律風險其實可以回歸為一個專題；亦即人工智慧的建構本身，其實已經明確存在受規範的可能性。並且相對而言，建立法秩序仍然是

法律最重要且歷史久遠的必要功能之一。因此，在有必要好好建置法規體系的概念下，重新思考人工智慧的定義、納入各方利害關係人的抉擇與影響力，並套入傳統法律概念中 Fiduciary Duty 的概念作為回應，對於回歸法律本身協助建立理想的社會環境，應存在有相當意義。

三、 研究方法

本文將主要採用文獻分析法（Literature Review）及比較法研究（Comparative Law），將國內外，現行多種切入人工智慧議題的治理概念，進行系統性研析；主要涵蓋 Balkin 所提出 The Fiduciary Model of Privacy 及相對應的辯論進行觀點交叉分析。並結合「質性內容分析（Qualitative Content Analysis）」，將所有可能參與人工智慧憲法的各類角色，包含私部門與企業、政府與公部門、民間社會團體、學術或技術社群等利害關係人的分析視角，藉由系統性分類，試論其中各方視角中重複出現的主題或模式、綜合其相異觀點，進行比較與討論。

四、 研究結果

首先，由（Employee Retirement Income Security Act of 1974, "ERISA"）所提出的概念可知，跳脫傳統形式的定義與既有法律關係與身份後，ERISA 於現代提出的法律概念更轉向功能判斷受托關係，將判斷是否實際行使對受益人利益或計畫資產具有實質影響的裁量權與控制力，建構 Fiduciary Duty[9]。美國最高法院在 Varity Corp. v. Howe 案例中亦採取此種功能性思考。

而人工智慧與使用者之間的關係，並非單純的資訊提供者。人工智慧不僅能夠取得、整理或分析特定使用者的資訊，更可能進一步根據這些資訊進行推理、提供建議，甚至代表使用者採取行動。換言之，人工智慧系統的影響力，無法僅以「間接影響」使用者而概括之，而是可能逐漸介入使用者的決策過程，並在資訊高度不對稱的情況下影響使用者的選擇與行為。這也對應從網際網路開始發展時，所謂數位平台即往往掌握使用者大量資訊，並透過其技術與資訊架構對使用者產生實質影響，因此傳統 fiduciary law 所關注的資訊優勢與權力不對稱，應可延伸至數位環境中的新型態關係[10]。

---

9 Dana Muir and Norman Stein. (2015). *Two hats, one head, no heart: The anatomy of the ERISA settlor/ fiduciary distinction*, 93 N.C. L. REV. 459, 475-476.

10 Jack M. Balkin, *Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation*, UC DAVIS L. REV. 51, 1149.

在人工智慧的情境中，這種關係更進一步受到系統開發者的設計與裁量所中介。使用者與人工智慧系統的每一次互動，表面上似乎是使用者與模型進行交流，但人工智慧系統的回應方式、記憶機制、資訊處理方式，以及何種行為受到鼓勵或限制，均受到開發者所建立的模型架構與治理規則影響。換言之，AI 對使用者所產生的影響能力，相當程度建立於使用者所自願交出的被控制權，並使開發者對系統行為具有持續且實質的控制力之上。

因此，如可透過實際行使及功能判準，而非形式、法律身分關係進行受託關係之判斷，則人工智慧系統之開發商所具有的功能性特徵將開啟進一步的法律介入空間。開發者不僅提供一項技術工具，而是透過系統設計決定如何理解使用者、如何回應使用者，以及在價值衝突出現時採取之判斷模式。可以確定的是，治理模式並非獨立互斥的關係，而是相輔相成，納入多方利害關係人的立場與理念，共同建構有效的人工智慧規範之重要基礎。

五、 問題討論

（一） 問題意識的起點：人工智慧治理與網路治理的並置討論

人工智慧本身伴隨的技術風險，面向多元且相互牽動，並與網際網路相輔交織發展。本文因此試圖以治理模式的脈絡進行分析。而核心關鍵在於，AI 並非憑空出現的新科技，而是在既有的網路基礎設施、資訊流通模式與平台經濟體系之中。因此 AI 所帶來的治理挑戰，實質上是對既有網路治理架構進行重構與反思[11]。本文亦指出，歷史經驗固然重要且具有參考價值，若僅依賴既有制度回應新問題，反而可能限縮治理想像力，忽略正在成形的新風險結構與全新權力配置，而陷入「典範負擔」[12]。在 AI 與網路高度交織的現實下，治理思維本身亦需同步調整。

（二） 生成式 AI 發展與網際網路所引領的治理議題

---

https://lawreview.law.ucdavis.edu/sites/g/files/dgvnsk15026/files/media/documents/51-3_Balkin.pdf (Last visited: Sep 20, 2026)

[11] Hadley, E., Blatecky, A. & Comfort, M. (2024). *Investigating algorithm review boards for organizational responsible artificial intelligence governance*. AI ETHICS 5, 2485–2495, 2489. https://doi.org/10.1007/s43681-024-00574-8. (Last visited: Sep 20, 2026)

[12] Garud, Raghu, and Michael A. Rappa. (1994). *A Socio-Cognitive Model of Technology Evolution: The Case of Cochlear Implants.* ORGAN. SCI. 5, no. 3 344–62, 348. http://www.jstor.org/stable/2635135. (Last visited: Sep 20, 2026)

生成式 AI 與高度個人化數位服務[13]，正因其風險與助益並存的現況，且正式形成全新社會常態，使相關治理的討論更顯迫切。此現象可以「高速公路增建，交通品質仍然不善」的關鍵比喻[14]，指出當制度改革、技術更新，往往同時刺激更高的使用強度與成本，而未必能夠真正改善整體品質。生成式 AI 所帶來的資訊環境變化，即呈現出上述類似悖論；大幅提升資訊取得與內容產出的速度與密度，看似降低使用門檻、提高生產效率，卻在另方面，將原本集中於媒體或平台端的查證與判讀成本，轉移至個別使用者身上。

在人類注意力（Attention Span）本身即屬於具備成本與排擠效應的資源前提下，使用者面對由系統主動生成的整合式內容[15]，必定花費大量時間與心力自行判斷其正確性、可信及完整度[16]，進而引發大量議題，涵蓋「AI 平台本身完整度需求，但訓練資料的正當性、平台所內涵 Input 層面連結人類基本隱私、著作、安全權利、操弄可能；以及 Output 方面廣泛在公共討論、資訊生態與社會行為所產生的直接與間接影響」，或許都可說是在人類有限性的基礎中展開。

歐洲研究報告指出[17]，生成式 AI 快速擴散，重塑新型態的人機與機器間互動模式，數位內容與廣告經濟結構重組，大型平台權力的現狀亦勢必遂調整；生成式 AI 不僅改變使用者取得資訊的方式，也動搖內容生產與流量的商業模式。當使用者得以直接透過 AI 取得答案，傳統內容產業與

---

[13] Crawford, K. (2021). *Introduction*. In The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence (pp. 1–21). YALE UNIV. PRESS.

[14] Duranton, Gilles, and Matthew A. Turner. (2011). *The Fundamental Law of Road Congestion: Evidence from US Cities*. AM ECON REV 101 (6): 2616–52, 2645. https://www.aeaweb.org/articles?id=10.1257/aer.101.6.2616. (Last visited: Sep 20, 2026)

[15] Robinson WS. (2014). *Philosophical challenges.* In: Frankish K, Ramsey WM, eds. The Cambridge Handbook of Artificial Intelligence. (pp. 64–86). CAMBRIDGE UNIVERSITY PRESS.

[16] Gyana Swain, *OpenAI Admits AI Hallucinations Are Mathematically Inevitable, Not Just Engineering Flaw*s, Computerworld (Sep18, 2025), https://www.computerworld.com/article/4059383/openai-admits-ai-hallucinations-are-mathematically-inevitable-not-just-engineering-flaws.html (Last visited: May 27, 2026).

[17] Sarah Sheil. (Jan, 2026). European Parliamentary Research Service. *10 issues to watch in 2026*, 10th ed., PE 782.587. European Parliament https://www.europarl.europa.eu/RegData/etudes/IDAN/2026/782587/EPRS_IDA(2026)782587_EN.pdf (Last visited: Sep 20, 2026)

依賴流量導向的數位廣告生態雖仰賴 AI 實行推廣，卻面臨衝擊，進而引發著作權、授權與商業模式的爭論。同時 AI 工具大幅縮短新科技達到關鍵使用規模的時間[18]（例如：Threads 等社群平台），使網路經濟的變化速度進一步加快，治理與法制回應的時間壓力也隨之上升。

（三） 平台角色轉變與 AI 法制回應的碎片化現象

傳統搜尋引擎模式下，平台主要扮演「資訊索引者」，並聚焦法律責任於事後咎責。然生成式 AI 中，平台成為主動生成內容的整合者，使平台責任、內容標示與使用者信賴關係面臨重塑，讓人類面臨難以預測的科技未來中[19]，不斷嘗試從前端破解問題，調整與試探。

當前各國對 AI 的法制回應即呈現高度碎片化；歐洲傾向密集與結構化的制度界定風險邊界[20]、美國更多仰賴市場競爭及政策引導與訴訟救濟、其他國家則嘗試基本法或原則以期先行建立治理框架。此處差異顯示出 AI 的治理逐漸引領世界走向「對話性」典範。

本文以為，在這種既日常又模糊的 AI 科技發展情境中，「資訊識讀能力」與「思辨能力」已然成為 AI 與網路時代不可或缺的能力。相應，這些思考或許應該進行轉向，AI 治理的重心在於，將整體法規遵循，前移納入組織內部的責任配置、使用準則與倫理架構[21]。

（四） 從網路治理經驗反思 AI 治理：組織、專業與多方參與

網路治理的發展經驗顯示，當問題高度跨國、跨技術且涉及多元利害關係人，單一國家或單一法規難以全面回應[22]。多方利害關係人模式的出現正是在此背景下逐步形成，其核心並非取代國家權威，而是透過對話與責任配置的制度

---

[18] Jemmy, Founder & CEO of JKL SEO. ChatGPT 統計數據彙整：GPT-4 的重要數據資料總覽 (Last modified: March 15, 2023) https://www.shopjkl.com/pages/dataanalysiscontest. (Last visited: Sep 20, 2026).

[19] 不如通訊科技在發展時，可以預測由 0%普及乃至 100%普及的電話設備。

[20] Artificial Intelligence Act, arts. 1–6, Regulation (EU) 2024/1689, EUR-Lex, https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng (Last visited: Sep 20, 2026).

[21] Mäntymäki, M., Minkkinen, M., Birkstedt, T. et al. (2022). *Defining organizational AI governance*. AI ETHICS 2, 603–609, 605. https://doi.org/10.1007/s43681-022-00143-x. (Last visited: Sep 20, 2026).

[22] Batool, A., Zowghi, D. & Bano, M. (2025). *AI governance: a systematic literature review*. AI ETHICS 5, 3265–3279, 3268. https://doi.org/10.1007/s43681-024-00653-w. (Last visited: Sep 20, 2026).

整合，建立可運作的治理秩序。AI 與網路治理中，法學訓練及律師的角色亦在轉變；不僅止於被動解釋法規，而在法律承擔的核心功能上，成為制度設計、風險管理與內控架構的一部分，包含分類技術、差異化規範對於治理工具的深化運用。企業實務中逐漸出現的 AI 使用守則、倫理準則與治理框架，即反映出在外部法制尚未定型前，組織內部治理所展現的高度彈性。

以新興社群平台 Threads 為例，即便其同溫層效應與社會影響備受質疑，但其快速成長亦揭示新技術所蘊含的巨大動能。因而更具建設性的理想在於，治理的目標在防範風險功能以外，是否將這些科技現實的力量轉化或納入成為引導政策、責任配置、並能創造公共價值的制度設計。

（五） 法制回應的分歧與共通

綜觀、盤點各國 AI 法制的不同路徑，無論是管控風險、原則導向或促進科技類型立法，面臨相似的核心難題在於，風險界定、責任歸屬、資料使用、創新效率與安全保障的悖論[23]。特別在個人資料保護與著作權領域，現行法律結構與 AI 技術發展之間存在顯著張力。而這些問題涉及資料生命週期、跨境流動、技術設計與制度選擇的整體調整。

本文結論以為，風險未必只能透過嚴密界定、責任未必僅能尋求單一歸屬、資料使用與創新發展亦非必然與安全對立。以台灣目前做法為例，現行「人工智慧基本法」採取風險分類（classification），而非進行歐盟 AI ACT 的分級管理（Risk-based regulation），呈現新型彈性治理取徑。

進一步言，日本將「互通性（interoperability）作為中心，結合文化及社會面向因素，不僅解決少子與勞力短少的問題，也將科技友善的在地價值融入，並強調事前協調的重要性，基本上，或許能將看似難以權衡（trade-off）的兩方天枰進行整合，在著重事前（ex ante）的共識時，不難想像支持科技目標發展，卻同時能夠對於責任透明、安全、風險同時完成全盤考量[24]。雖然不如歐洲以引領性指導原則（The

---

[23] Floridi, L. (2021). *The European Legislation on AI: a Brief Analysis of its Philosophical Approach.* PHILOS. TECHNOL. 34, 215–222, 218. https://doi.org/10.1007/s13347-021-00460-9 (Last visited: Sep 20, 2026).

[24] Habuka, H. (March 21, 2025). *Unpacking Japan's AI policy with Hiroki Habuka.* Center for Strategic and International Studies, https://www.csis.org/analysis/unpacking-japans-ai-policy-hiroki-habuka (Last visited: Sep 20, 2026).

Brussel Effect）為中心塑造全球標準、或不同於美國強力發達的 AI Superpower 及訴訟救濟生態，但既然這是一個全新生態場域，在 AI 與網路科技已成為地緣政治資產的時代[25]，台灣或是任何其他主體可以做到的是，藉由劃定明確的核心價值，則可以期待開創未來科技與人類社會共榮的願景[26]

---

[25] David U. Socol de la Osa. Supra note 5.

[26] Habuka, H. (Oct 9, 2025). *Japan's Agile AI Governance in Action: Fostering a Global Nexus Through Pluralistic Interoperability*, Center for Strategic and International Studies. *https://www.csis.org/analysis/japans-agile-ai-governance-action-fostering-global-nexus-through-pluralistic* (Last visited: Sep 20, 2026).