算法治理觀察 下篇

算法治理遇挑戰 確保技術安全合規 避免過度監管 倡建技術聯盟促行業自律 創新與監管須平衡

●國家網信辦等四部委提出，要常態化開展算法服務安全風險監測防範工作，提升算法常態化治理水平。圖為觀眾在參觀網絡數據分析大屏幕。 資料圖片

數字社會正在到來，算法治理是數字時代公共管理的全新命題，也是政府治理的重大挑戰之一。經過多次專項治理後，國家網信辦等四部委在11月啟動的「清朗・網絡平台算法典型問題治理」專項行動中強調，要推動長效治理，常態化開展算法服務安全風險監測防範工作，提升算法常態化治理水平。同濟大學經濟與管理學院副教授鄧弘林在接受香港文匯報採訪時表示，「算法的迅速迭代和自主學習能力，使得其創新速度遠遠超出現有監管手段和法律法規的更新速度，監管的滯後性和算法的技術壁壘之間的矛盾，構成了常態化治理的重大挑戰。」多位受訪專家認為，人工智能算法事關社會經濟發展，如何平衡算法技術的創新發展和有效監管是算法治理的關鍵問題。

●香港文匯報記者 劉凝哲、蘇雨潤 北京報道

鄧弘林認為，在常態化治理過程中，如何確保平台在追求商業利益的同時，能夠履行應有的社會責任並遵循倫理規範，是一個亟待解決的問題。許多平台算法的設計初衷是為了提高流量和用戶黏性，這往往導致算法優化以商業利益為主導，忽視了潛在的社會風險，例如算法可能帶來的歧視、偏見或誤導消費者等問題。如何平衡平台的商業需求和社會責任，確保算法的公正性與透明度，是治理中的一大難題。同時，增強公眾對算法治理的認同與信任，是推動治理常態化的重要前提。如果公眾對治理機制缺乏信任，可能會導致相關政策和措施的執行困難，甚至產生抵觸情緒。

「我認為目前最大的挑戰在於如何平衡算法的創新發展與有效監管。」中國傳媒大學新聞學院教授詹騫表示，一方面，算法作為現代科學技術的重要成果，已經深刻嵌入社會發展的各個領域，其創新應用能夠推動市場環境更透明、信息更易於自由流動，降低搜尋成本和准入門檻，同時促進技術突破和效率提升。另一方面，算法的複雜性和隱蔽性也給監管帶來了巨大難度，如何確保算法在發揮積極作用的同時，不侵犯用戶權益、不產生算法歧視、不助力壟斷等問題是當前治理面臨的重要課題。

兩年半 中央四次出手治理AI算法

2024年11月
「清朗・網絡平台算法典型問題治理」專項行動
重點整治同質化推送營造「信息繭房」、違規操縱干預榜單炒作熱點、盲目追求利益侵害新就業形態勞動者權益、利用算法實施大數據「殺熟」、算法向上向善服務缺失侵害用戶合法權益等重點問題，督促企業深入對照自查整改，進一步提升算法安全能力。

2023年12月
「清朗·整治短視頻信息內容導向不良問題」專項行動
優化推薦機制，着力解決短視頻平台算法價值導向存在偏差、優質短視頻呈現不足等問題。優化流量分配機制，防止「重指標輕質量」，片面以點讚率、轉發率等量化指標作為流量分配依據。

2023年9月
「清朗・生活服務類平台信息內容整治」專項行動
聚焦為線下違法活動引流、搜索環節呈現違法信息、發布違規營銷信息、組織操縱刷分控評、重點環節推薦低俗不良信息、傳播網絡迷信信息、散布炫富拜金、暴飲暴食信息等7類突出問題，集中排查整治。

2022年4月
「清朗・2022年算法綜合治理」專項行動
深入排查整改互聯網企業平台算法安全問題，評估算法安全能力，重點檢查具有較強輿論屬性或社會動員能力的大型網站、平台及產品，督促企業利用算法加大正能量傳播、處置違法和不良信息、整治算法濫用亂象、積極開展算法備案，推動算法綜合治理工作的常態化和規範化。

監管體系應動態靈活 有行業差異化

「為了在推動技術創新的同時避免過度監管對發展的抑制，可以採用漸進式的分階段監管模式」，鄧弘林提出，在技術初期，應該盡量給予創新較大的空間，支持研發和實驗，以鼓勵技術突破。在技術逐步成熟並廣泛應用後，可以通過對應用場景的深入評估，逐步建立起更加嚴格的監管機制。

鄧弘林建議，不同領域和應用場景下，算法所面臨的風險和挑戰不同，因此監管框架應具有靈活性和差異化設計。比如，在金融、醫療等高度敏感的行業，算法應用的監管應更加嚴格，重點關注算法的透明性、可解釋性、數據隱私保護以及避免歧視和偏見等問題。而對於娛樂或社交平台的推薦算法，監管的重點則可以放在防止信息誤導、避免算法陷入「回音室效應」（即一些相近觀點在相對封閉的環境中不斷重複）以及防範過度依賴算法等方面。通過靈活設計監管框架，不僅能確保創新在不同行業中的適用性，還能在關鍵領域有效管控潛在的社會風險。

此外，隨着技術進步和應用場景的不斷變化，「建立一個動態的監管體系至關重要。」鄧弘林表示，監管機構可要求企業定期進行算法的風險評估，檢查其在實際應用中的表現，並評估是否帶來新的社會風險。法規和監管機制也應具備靈活性。動態風險評估機制應結合社會變革、技術發展和用戶反饋，及時調整監管措施。

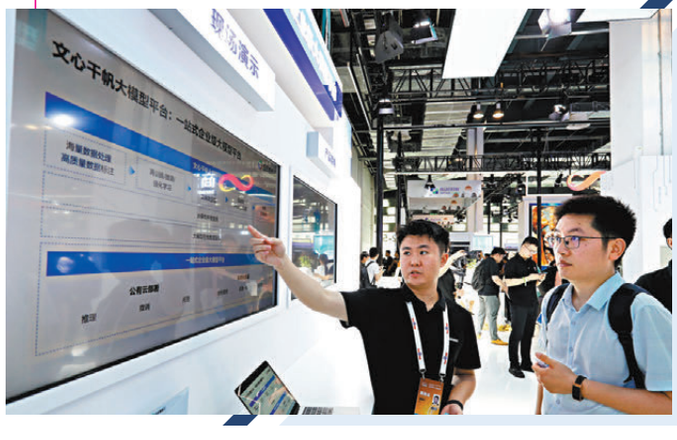
行業自律在算法治理中也起到至關重要的作用，在鄧弘林看來，可以鼓勵建立行業協會和技術聯盟，推動行業內企業共同制定技術標準和倫理守則，從而形成自我約束機制。

跨部門合作 避免監管競合

「在人工智能領域，監管可採取『審慎包容』的態度，既確保技術的安全和合規，又要避免過度監管成為發展的『緊箍咒』。」內地知名人工智能公司中科聞歌人工智能專家表示，由於人工智能技術的跨領域特性，需要多個部門協同合作，避免監管競合。可以通過跨部門合作，形成合力，共同為場景創新提供制度供給，促進人工智能創新發展與監管規範相協調。

具體措施方面，中科聞歌人工智能專家認為，可建立一套涵蓋治理目標、治理主體、治理對象、治理手段和治理模式的體系化綜合治理框架。在治理目標上實現算法可問責與算法經濟高質量發展，在治理主體上通過部際聯席會議制度形成算法治理合力，在治理對象上拓寬算法治理的範圍，在治理手段上完善技術治理制度，在治理模式上優化多元主體參與的協同共治模式。此外，還要鼓勵技術創新，制定算法倫理標準和認證機制，對AI算法生命周期中的風險進行管控。同時，引入獨立專業的第三方機構，搭建起成熟統一的評估標準，對算法的合理性進行審查。此外，有必要在監管中嵌入技術工具，以技術手段賦能監管，打造與人工智能體模塊化治理框架相適配的智慧化治理工具，實現對特定高風險場景的精準化治理。

●人工智能技術的跨領域特性，需要多個部門協同合作。圖為工作人員（左）在向參觀者介紹百度文心千帆大模型平台。 資料圖片



●專家表示算法作為現代科學技術的重要成果，已經深刻嵌入社會發展的各個領域。圖為選手在工業大數據算法項目比賽中操作。 資料圖片

推動算法透明化 加快數據資源立法

數據驅動的算法時代，機器在圖像識別、語言處理等領域取得顯著進步，但在此過程中產生的隱私洩露、算法歧視、信息繭房等問題日益凸顯。中國傳媒大學新聞學院副教授李建剛表示，構建完善的數字社會司法體系已是當務之急。

李建剛指出，政府通過法規規範企業和平台行為是算法治理的重要基礎，但要真正觸及問題核心，需要解決兩個更深層次的挑戰：一是平台的盈利驅動，二是流量分配權力的不對稱性。當前，許多平台的算法設計以點擊率和用戶停留時長為優化目標，導致低俗、煽動性內容氾濫，而商業內容往往佔據大部分流量資源。同時，平台掌握着流量分配的核心權力，控制着內容曝光的優先級，這不僅加劇了信息不平等，還可能對社會認知和公共輿論產生外溢效應。

「為破解這些問題，政府需要採取更深層次的干預措施，突破傳統監管的局限。」李建剛建議，首先要推動算法透明化，要求平台公開核心算法邏輯，尤其是流量分配的主要參數和機制，並制定透明性標準，明確哪些內容應該優先推薦，保障多樣性和公共價值內容的基本曝光。其次，應在算法設計中引入「公共價值」目標，通過調整權重確保公共新聞和公益內容得到合理的推薦，同時推動平台優化平衡機制，在商業利益與社會責任之間找到動態平衡點。

立法與執法雙管齊下

在立法與執法方面，北京資深法律人李玉斌認為，為應對數字社會帶來的挑戰，需要制定一系列針對性強、操作性高的法律法規。在立法上，應高度重視數據要素的所有權、使用權、監管權以及信息保護和數據安全等方面的立法工作，確保各類數據資源合法安全使用。在執法上，要建立健全數字社會監管體系，加強平台監管，確保其合法合規運營。同時，執法機構需加強對於算法的學習，具備高技術和專業素養，及時發現並處理違法行為。

李玉斌還建議，政府可以通過建立「合作沙盒機制」與平台共同測試算法對流量分配的影響，確保算法的運行符合法律法規和社會倫理要求，強化第三方監督，通過獨立的算法審計機構對平台的流量分配機制進行定期審查，並向公眾公開結果，增強治理的透明度和權威性。

展開安全評估 實施備案管理

中國科學院院士張鈸早前提出疑問，「機器人會不會成為我們的主人？如果機器的智能超過我們，我們就失去對它的控制，對於這樣的機器人要加以限制，加以治理。」針對如何避免算法學習的不可預知性和不可控性，中科聞歌人工智能專家表示，從技術層面，可以建立和加強算法影響性評估機制（Algorithmic Impact Assessment，AIA），以及設置採購規則來強化公共部門所使用算法的透明度。

對於具有輿論屬性或社會動員能力的生成式人工智能服務，專家建議開展安全評估，並履行算法備案和變更、註銷備案手續，以應對算法「黑箱」問題。有關主管部門對生成式人工智能服務開展監督檢查時，提供者應依法配合，對訓練數據來源、規模、類型、標注規則、算法機制機理等予以說明，並提供必要的技術、數據和文檔等支持和協助。

籲限制算法自主決策範圍

同濟大學經濟與管理學院副教授鄧弘林認為，為避免算法自我進化的過程中出現不可預測，導致算法行為的失控或出現不良後果，需要對算法的學習過程施加適當的限制，尤其是在金融、醫療、公共安全等關鍵領域。例如，可以限制算法自主決策的範圍，確保其始終在人類可控的框架內運行。對於高風險領域，可以實施「人機協作」模式，即在重要決策環節引入人工審核或干預，確保算法的決策符合社會的長遠利益。

從國際經驗看，西方國家亦積極開展算法治理。美國拜登政府發布《人工智能權利法案藍圖》、《關於安全、可靠、可信地開發和使用人工智能的行政令》；通過召集人工智能頭部企業座談並敦促其發布《人工智能自願承諾》，推動行業自律；在企業界，包括特斯拉總裁馬斯克在內的1,000多位美國科技精英聯名發聲，呼籲暫停訓練比GPT-4更強大的AI系統至少6個月，以確保人類能夠緊跟人工智能的發展步伐。在國際上，美國主動參與並推動一系列合作機制與倡議，力圖在全球人工智能治理中佔據主導地位。歐盟《數字服務法案》則在2022年生效，提出採取算法問責和透明度審計等措施，要求在線平台公開算法參數，提高透明度；還成立歐洲算法透明度中心（ECAT），提供科學和技術支持。