七部門公布生成式AI管理暫行辦法

採包容審慎分類分級監管 明確法律責任

香港文匯報訊（記者 王玨 北京報道）近日，國家網信辦聯合六部委公布《生成式人工智能服務管理暫行辦法》（下稱《辦法》），旨在促進生成式人工智能健康發展和規範應用，維護國家安全和社會公共利益，保護公民、法人和其他組織的合法權益。《辦法》要求，採取有效措施鼓勵生成式人工智能創新發展，明確了提供和使用生成式人工智能服務總體要求，對生成式人工智能服務實行包容審慎和分類分級監管。此外，《辦法》還明確網絡安全法等系列法律為有關主管部門的參考罰則。

國家網信辦聯合國家發改委、教育部、科技部、工業和信息化部、公安部、國家廣電總局公布的該《辦法》，自2023年8月15日起施行。國家網信辦有關負責人指出，近年來，生成式人工智能技術快速發展，為經濟社會發展帶來新機遇的同時，也產生了傳播虛假信息、侵害個人信息權益、數據安全和偏見歧視等問題，如何統籌生成式人工智能發展和安全引起各方關注。出台《辦法》，既是促進生成式人工智能健康發展的重要要求，也是防範生成式人工智能服務風險的現實需要。

防範未成年人用戶過度沉迷

《辦法》明確，生成式人工智能技術，是指具有文本、圖片、音頻、視頻等內容生成能力的模型及相關技術。生成式人工智能服務提供者，是指利用生成式人工智能技術提供生成式人工智能服務（包括通過提供可編程接口等方式提供生成式人工智能服務）的組織、個人。生成式人工智能服務使用者，是指使用生成式人工智能服務生成內容的組織、個人。

《辦法》適用於利用生成式人工智能技術向中華人民共和國境內公眾提供生成文本、圖片、音頻、視頻等內容的服務。《辦法》提出國家堅持發展和安全並重、促進創新和依法治理相結合的原則，採取有效措施鼓勵生成式人工智能創新發展，對生成式人工智能服務實行包容審慎和分類分級監管，明確了提供和使用生成式人工智能服務總體要求。提出了促進生成式人工智能技術發展的具體措施，明確了訓練數據處理活動和數據標註等要求。規定了生成式人工智能服務規範，明確生成式人工智能服務提供者應當採取有效措施防範未成年人用戶過度依賴或者沉迷生成式人工智能服務，按照《互聯網信息服務深度合成管理規定》對圖片、視頻等生成內容進行標識，發現違法內容應當及時採取處置措施等。

◆國家網信辦聯合六部委公布《生成式人工智能服務管理暫行辦法》，旨在促進生成式人工智能健康發展和規範應用。圖為一家廠商的「人工智能生成圖像」。 新華社

需政企社會網民多方參與

此外，《辦法》還規定了安全評估、算法備案、投訴舉報等制度，明確了法律責任。明確提供者違反本辦法規定的，由有關主管部門依照《中華人民共和國網絡安全法》、《中華人民共和國數據安全法》、《中華人民共和國個人信息保護法》、《中華人民共和國科學技術進步法》等法律、行政法規的規定予以處罰；法律、行政法規沒有規定的，由有關主管部門依據職責予以警告、通報批評，責令限期改正；拒不改正或者情節嚴重的，責令暫停提供相關服務。

國家網信辦有關負責人指出，生成式人工智能服務的發展與治理需要政府、企業、社會、網民等多方參與，共同促進生成式人工智能健康發展，讓生成式人工智能技術更好地造福人民。

《辦法》適用範圍

- 利用生成式人工智能技術向中華人民共和國境內公眾提供生成文本、圖片、音頻、視頻等內容的服務，適用本辦法。
- 對利用生成式人工智能服務從事新聞出版、影視製作、文藝創作等活動另有規定的，從其規定。
- 行業組織、企業、教育和科研機構、公共文化機構、有關專業機構等研發、應用生成式人工智能技術，未向境內公眾提供生成式人工智能服務的，不適用本辦法的規定。

整理：香港文匯報記者 王玨

《辦法》要點(部分)

1、堅持社會主義核心價值觀，不得生成煽動顛覆國家政權、推翻社會主義制度，危害國家安全和利益、損害國家形象，煽動分裂國家、破壞國家統一和社會穩定，宣揚恐怖主義、極端主義，宣揚民族仇恨、民族歧視，暴力、淫穢色情，以及虛假有害信息等法律、行政法規禁止的內容；

2、在算法設計、訓練數據選擇、模型生成和優化、提供服務等過程中，採取有效措施防止產生民族、信仰、國別、地域、性別、年齡、職業、健康等歧視；

3、尊重知識產權、商業道德，保守商業秘密，不得利用算法、數據、平台等優勢，實施壟斷和不正當競爭行為；

4、尊重他人合法權益，不得危害他人身心健康，不得侵害他人肖像權、名譽權、榮譽權、隱私權和個人信息權益；

5、使用具有合法來源的數據和基礎模型；

6、涉及知識產權的，不得侵害他人依法享有的知識產權；

7、涉及個人信息的，應當取得個人同意或者符合法律、行政法規規定的其他情形；

8、參與生成式人工智能服務安全評估和監督檢查的相關機構和人員對在履行職責中知悉的國家秘密、商業秘密、個人隱私和個人信息應當依法予以保密，不得洩露或者非法向他人提供；

9、對來源於中華人民共和國境外向境內提供生成式人工智能服務不符合法律、行政法規和本辦法規定的，國家網信部門應當通知有關機構採取技術措施和其他必要措施予以處置。

整理：香港文匯報記者 王玨

開啟中國式AIGC監管之路 劃定安全底線

專家解讀

北京專家分析指出，面對由自然語言模型ChatGPT掀起的生成式人工智能（AIGC）浪潮，中國適時完善中國人工智能治理體系的框架，開啓了中國式AIGC監管之路。可以看到，中國對於該領域的監管和治理原則，在積極引導產業發展的同時，也注重劃定安全底線，突出國家安全和個人權益保護，這有助於降低新技術帶來風險和衝擊，亦是因應當前國內外形勢的現實需求。

根據經濟合作與發展組織（OECD）匯總的信息，目前有數十個國家以及歐盟針對人工智能發展提出了戰略規劃，有的已推出或正在研究指導意見和監管措施。其中歐盟在2021年就提出了《人工智能法案》，強調要通過建立法規監管體系保證人工智能發展安全、符合道德規範且值得信賴。美國聯邦貿易委員會5月也表示，該機構致力於利用現有法律來控制人工智能的風險。

突出國家安全和個人權益保護

網絡安全學者、清博研究院副研究員葉晨宇對香港文匯報分析稱，基於不同的發展環境，在「技術—社會—法律」的體系下，各國均已開始嘗試法律規制人工智能的手段。中國對該領域的監管和治理原則，在積極引導產業發展的同時，也突出國家安全和個人權益保護，體現了中國特色。

她注意到，今年4月11日，國家網信辦就《生成式人工智能服務管理辦法》公開徵求意見。而此次發布的《辦法》中，新增多處對於AIGC的相關鼓勵，力促AIGC發揮更大的社會和商業價值，這將爲中國人工智能產業的發展注入強大動力，進一步提升中國在人工智能領域的國際競爭力。

與此同時，新技術的發展可能被惡意利用，成爲實施犯罪活動的基礎設施，如AI換臉視頻詐騙，甚至破壞國家安全等。她指出，此次發布的《辦法》，對AIGC服務提出了多項規制，強調AIGC產品提供者的責任、突出堅持社會主義核心價值觀、維護國家統一和社會穩定、保護個人信息、明確向監管部門備案和申報安全評估的硬性要求，以及多次重申要從數據源開始確保「生成內容」的眞實準確等，有助於完善中國人工智能治理體系的框架，同時應對生成式人工智能帶來的風險與衝擊。

◆香港文匯報記者 王玨 北京報道

涉個人信息需取得個人同意

香港文匯報訊（記者 任芳頡 北京報道）《辦法》明確，生成式人工智能服務提供者應當依法開展預訓練、優化訓練等訓練數據處理活動，使用具有合法來源的數據和基礎模型；涉及知識產權的，不得侵害他人依法享有的知識產權；涉及個人信息的，應當取得個人同意或者符合法律、行政法規規定的其他情形。

「AI換臉」不是想換就換

人工AI問世以來，其技術的便捷性使之在使用上越來越遊走在灰色地帶。例如，早前一名女網友被人利用AI技術「一鍵脫衣」，使一張正常的地鐵照被處理成涉黃照，進而給該網友帶來諸多困擾。再比如自稱「渾元形意太極門掌門人」的「馬保國」近來一直遭遇網民惡搞，有好事者通過AI技術將其面部移接至一些搞笑視頻，令到當事人不堪其擾，卻申訴無門。「AI換臉」詐騙，濟南一男子因視頻通話7秒鐘被騙30萬元；超百部「明星、網紅換臉」色情視頻僅售幾十元……越來越多網民擔憂下一個AI技術的受害者可能就是自己。

「『一鍵換臉』技術的應用常常未經原創內容的版權所有者同意，就私自改動其作品，涉嫌侵犯了他人的知識產權。」聯合國國際法委員會顧問、瑞中法律協會執行理事張天澤認為，新辦法的實施將對此類行為進行嚴格規範，任何未經版權所有者許可的「換臉」行為都將被視為非法。

AI生成產品需明確標註

香港文匯報訊（記者 任芳頡 北京報道）《辦法》還規定，向中國境內公眾提供生成文本、圖片、音頻、視頻等內容的服務需進行標註，防止出現虛假信息。

濫用AI增加鑒別眞偽難度

聯合國國際法委員會顧問、瑞中法律協會執行理事張天澤表示，AI技術能夠生成逼真的文本、圖片、音頻和視頻內容，提供更為豐富的互動體驗。然而，如果這些內容被惡意使用，可能會誤導公眾，影響社會穩定。比如，AI虛擬主播可以全天候工作，用逼真的聲音和面部表情報告新聞，給人一種真人主播正在播報新聞的體驗。

然而，隨着該技術的普及，出現濫用情況，例如，今年3月，一組「特朗普被捕」的「照片」在社交平台流傳開來。人物表情的真實、場景甚至鏡頭景深等細節，令諸多見慣類似場景的新聞攝影師也難辨真偽。而這組照片則是AI根據人類指令合成的。這也意味着人工智能生成圖文或影片增加了網民識別信息真偽的成本和難度，更為社會的安定帶來不小的隱憂。

◆7月7日，觀眾參觀主題為「交響」的AIGC藝術展。 新華社

「例如通過虛擬主播發布虛假或誤導性的信息，嚴重影響了公眾對信息的正確理解。新規出台後，所有AI生成的內容都必須明確標註，公眾知曉後可以更加謹慎地對待，有效提高信息鑒別能力。」

張天澤分析，新規實施後，將對向公眾提供的AI生成內容進行嚴格的監管，違規行為將被嚴厲打擊。之前存在不規範行為將被清理，行業的整體合規性和公信力將得到提升。與此同時，企業可能需要投入更多資源來獲取和管理合法的數據，尋求新的技術或業務模式來應對，這些額外的投入可能會對企業的運營成本產生一定的影響，將導致行業的優勝劣汰，推動技術創新和商業創新。

關注訓練數據 防歧視免誤用

香港文匯報訊（記者 任芳頡 北京報道）人工智能是雙刃劍，很多科技界大佬對人工智能無序發展持悲觀態度。由於AI訓練成果取決於學習數據，新公布的《辦法》明確規定，提供和使用生成式人工智能服務，應當遵守法律、行政法規，尊重社會公德和倫理道德。包括但不限於禁止生成煽顛國家政權、危害國家安全、宣揚恐怖極端主義、民族仇恨的內容，訓練數據需有效採取防止產生民族、信仰、國別、地域、性別、年齡、職業、健康等歧視。

以人為本 適時審視系統

「目前一些私企和軍事部門已經開始開發利用人工智能負面效應功能，人工智能不僅僅是技術進步，也意味着未來可能在戰場上替人類做出災難性決定。」聯合國日內瓦辦事處總幹事辦公室主任David Chikvaidze近日指出，從技術層面來講，人工智能還需要進行更多的技術改進，應該更加關心人工智能是否優於人類智能，以及因人性特徵可能引發的誤用問題。

聯合國國際法委員會顧問、瑞中法律協會執行理事張天澤近來參加多場全球性關於人工智能討論的峰會，他對香港文匯報表示，生成式人工智能服務訓練數據必須依法依規，要遵循反對歧視的準則，包括消除歧視。

「例如在招聘過程中，AI可以提高效率，幫助企業從大量的簡歷中篩選出最合適的候選人。然而，如果不正確地使用，可能會導致對某些群體的歧視。」張天澤舉例稱，美國電商巨頭Amazon曾使用一款AI招聘工具來自動篩選簡歷，以幫助招聘團隊找到最佳候選人。然而，該工具主要使用過去的招聘數據進行訓練，這些數據中男性應聘者佔主導地位，使得該工具傾向於推薦男性候選人。張天澤稱，根據新規，公司需要重新調整他們的AI招聘系統，以確保不論性別、年齡、地域等，每一個候選人都能得到公平的評估。