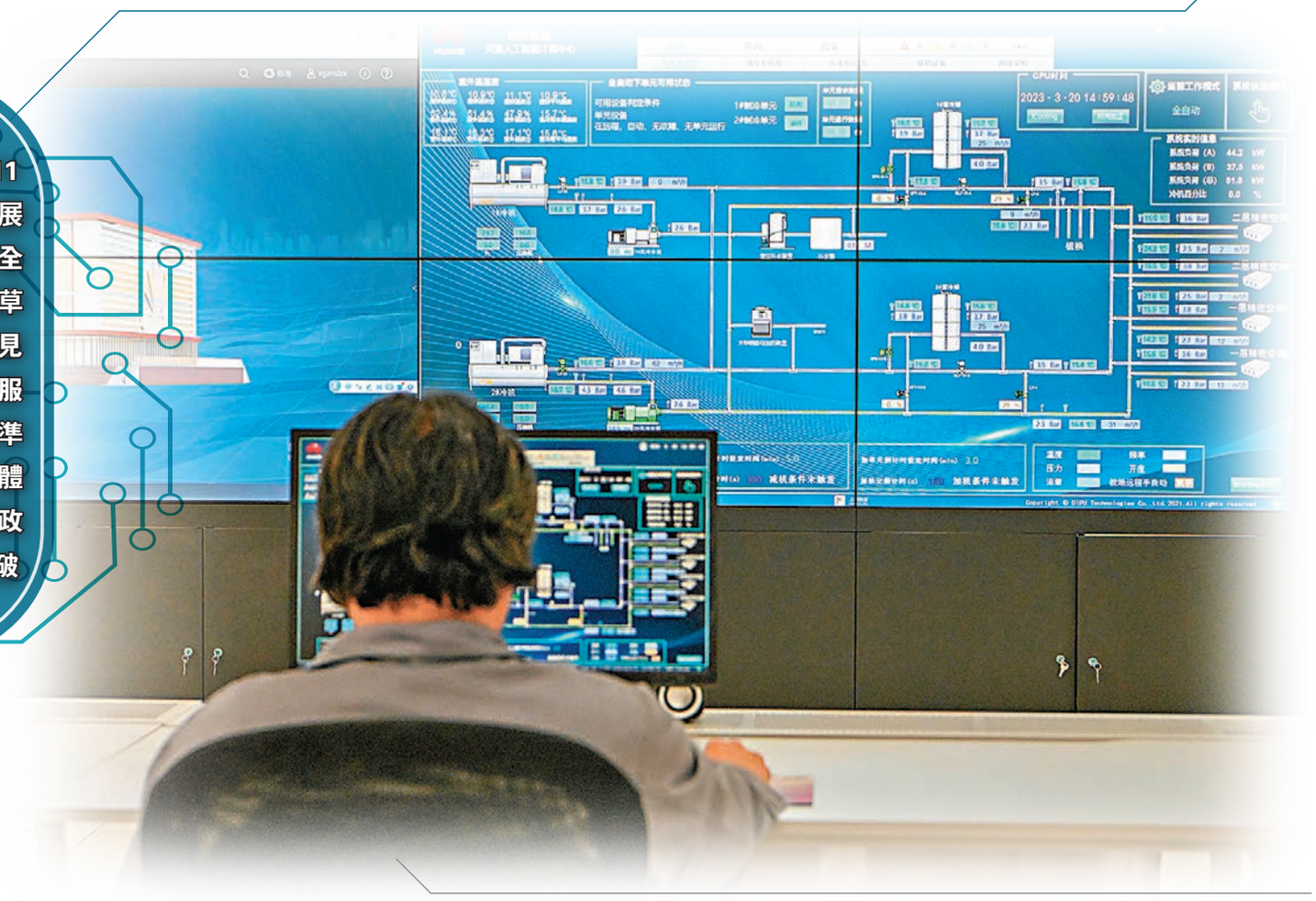
# 網信辦發布《生成式人工智能服務管理辦法(徵求意見稿)》

**香港文匯報訊（記者 張帥 北京報道）國家互聯網信息辦公室官網4月11日消息，爲促進生成式人工智能技術健康發展和規範應用，根據《中華人民共和國網絡安全法》等法律法規，國家互聯網信息辦公室起草了《生成式人工智能服務管理辦法（徵求意見稿）》，其中提出生成式人工智能產品提供服務前需申報安全評估，在內容上應當眞實準確，採取措施防止生成虛假信息，並且應當體現社會主義核心價值觀，不得含有顛覆國家政權、推翻社會主義制度，煽動分裂國家、破壞國家統一等內容。**



◆國家網信辦起草了《生成式人工智能服務管理辦法（徵求意見稿）》，其中提出生成式人工智能產品提供服務前需申報安全評估，在內容上應當眞實準確，採取措施防止生成虛假信息。圖為今年3月在天津市人工智能計算中心中控室，技術人員在監控設備運行情況。 資料圖片

## 生成式人工智能應防止虛假信息

生成式人工智能，是指基於算法、模型、規則生成文本、圖片、聲音、視頻、代碼等內容的技術。隨着OpenAI公司聊天機器人ChatGPT和百度「文心一言」流行，生成式人工智能成為全球焦點，相關監管也逐漸提上日程。

今次國家互聯網信息辦公室發布的《徵求意見稿》明確要求，利用生成式人工智能產品向公眾提供服務前，應當按照《具有輿論屬性或社會動員能力的互聯網信息服務安全評估規定》向國家網信部門申報安全評估，並按照《互聯網信息服務算法推薦管理規定》履行算法備案和變更、註銷備案手續。

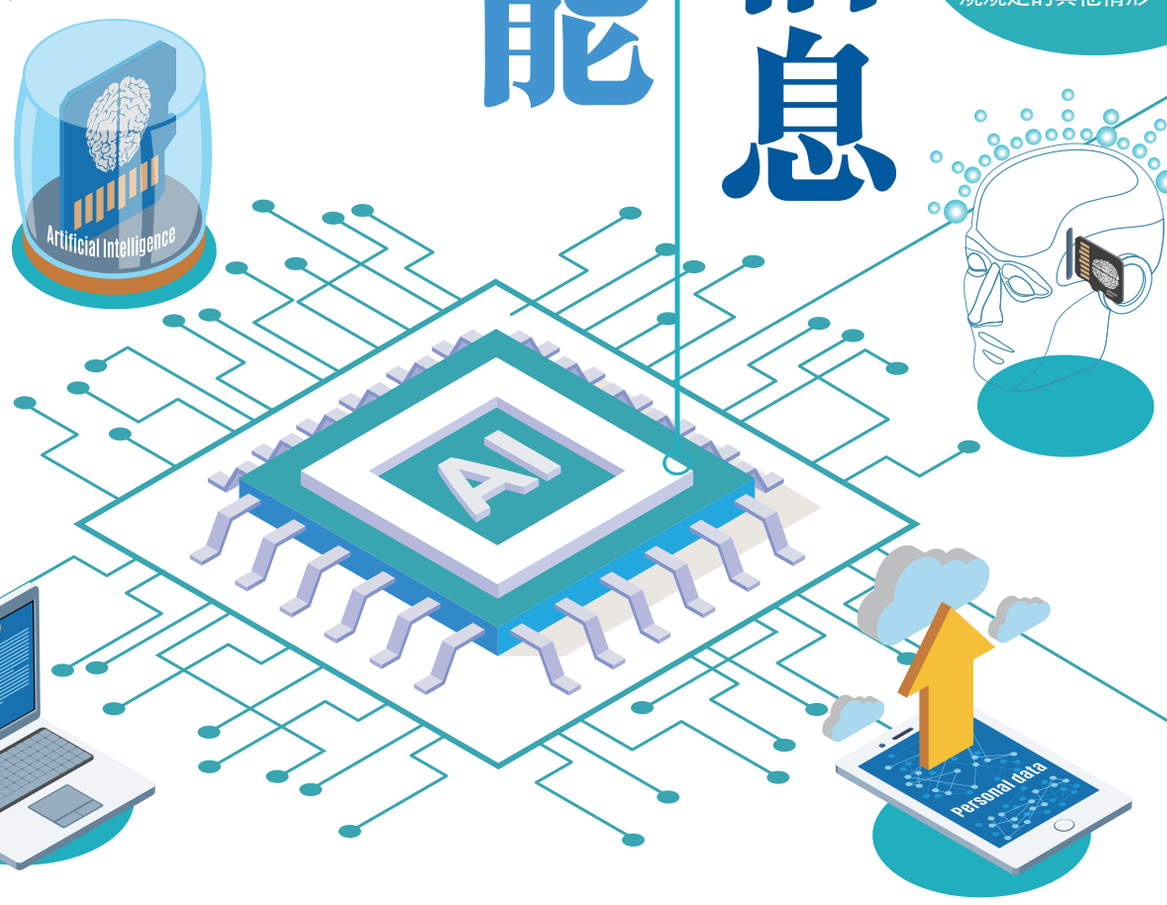
### 提出五方面具體要求

《徵求意見稿》指出，國家支持人工智能算法、框架等基礎技術的自主創新、推廣應用、國際合作，鼓勵優先採用安全可信的軟件、工具、計算和數據資源。

提供生成式人工智能產品或服務應當遵守法律法規的要求，尊重社會公德、公序良俗，符合以下五方面要求。**一是**利用生成式人工智能生成的內容應當體現社會主義核心價值觀，不得含有顛覆國家政權、推翻社會主義制度，煽動分裂國家、破壞國家統一，宣揚恐怖主義、極端主義，宣揚民族仇恨、民族歧視，暴力、淫穢色情信息，虛假信息，以及可能擾亂經濟秩序和社會秩序的內容。**二是**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，採取措施防止出現種族、民族、信仰、國別、地域、性別、年齡、職業等歧視。**三是**尊重知識產權、商業道德，不得利用算法、數據、平台等優勢實施不公平競爭。**四是**利用生成式人工智能生成的內容應當真實準確，採取措施防止生成虛假信息。**五是**尊重他人合法利益，防止傷害他人身心健康，損害肖像權、名譽權和個人隱私，侵犯知識產權。禁止非法獲取、披露、利用個人信息和隱私、商業秘密。

### 不得非法留存用戶輸入信息

此外，《徵求意見稿》提出，提供者在提供服務過程中，對用戶的輸入信息和使用紀錄承擔保護義務。不得非法留存能夠推斷出用戶身份的輸入信息，不得根據用戶輸入信息和使用情況進行畫像，不得向他人提供用戶輸入信息。提供者不得根據用戶的種族、國別、性別等進行帶有歧視性的內容生成。



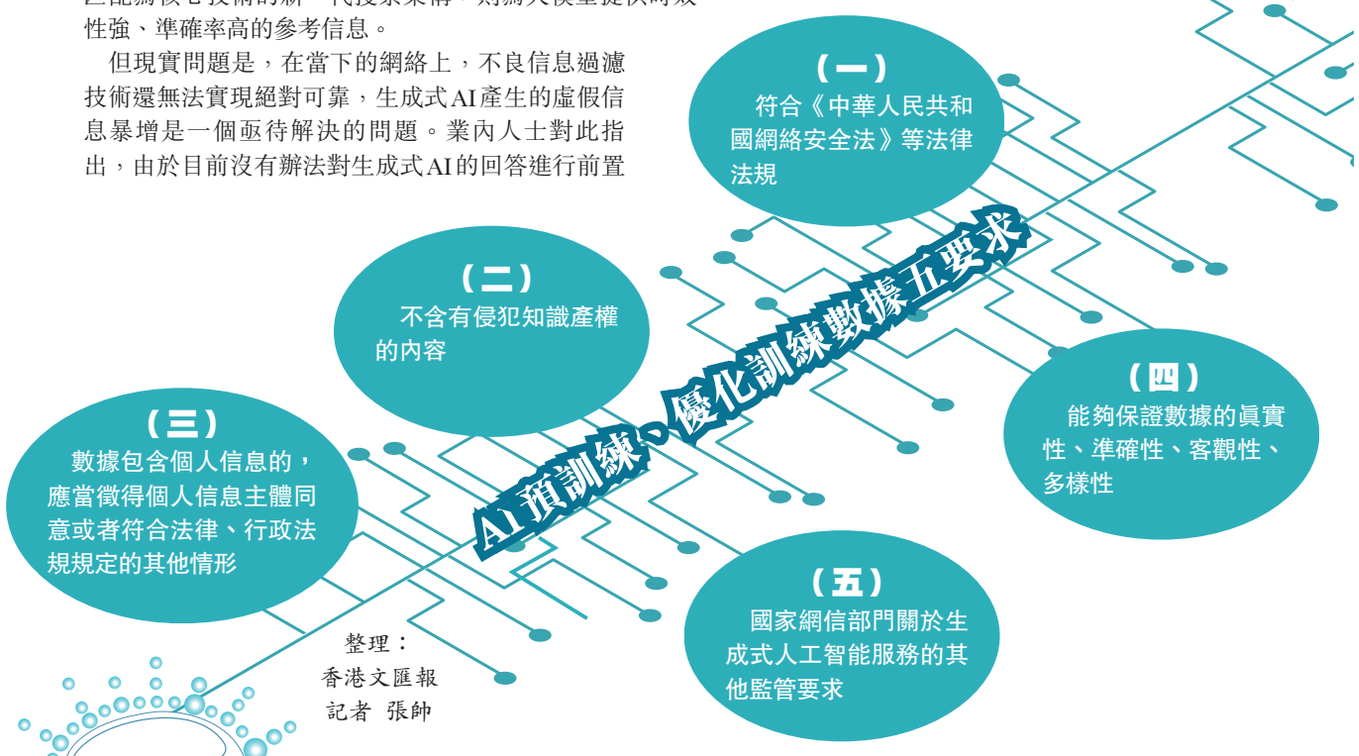

## AI不良信息暴增 監管應劃定紅線

**香港文匯報訊（記者 張帥 北京報道）**業內人士對香港文匯報記者介紹，海量數據是生成式AI的訓練基礎，通過每天接受數十億次用戶的搜索請求，生成式AI能夠基於龐大、高效的數據池快速地學習和改進。而用戶的反饋越多，人工智能生成內容的準確性就會越來越高，效果會越來越好。因此，隨着應用逐漸普及，監管也要跟上步伐，其中重要的是要劃定好應用紅線。

生成式AI要對海量數據進行預訓練和優化訓練。以百度文心一言大模型為例，其訓練數據包括萬億級網頁數據、數十億的搜索數據和圖片數據、百億級的語音日均調用數據，以及5,500億事實的知識圖譜等。不僅可以通過「知識內化」，從大規模知識和無標註數據中利用知識構造訓練數據，也能夠通過「知識外用」，引入外部多源異構知識，做知識推理、提示構建等。同時，以語義理解與語義匹配為核心技術的新一代搜索架構，則為大模型提供時效性強、準確率高的參考信息。

但現實問題是，在當下的網絡上，不良信息過濾技術還無法實現絕對可靠，生成式AI產生的虛假信息暴增是一個亟待解決的問題。業內人士對此指出，由於目前沒有辦法對生成式AI的回答進行前置性審核，在海量回答中，包含有嚴重錯誤的答案很容易「逃逸」。在監管上，對此並非無計可施，如可以在技術上開發出能夠辨別AI生成內容的新技術，這樣幫助信息獲取者了解其面對的內容是否是AI合成。其次，面向人工智能技術的開發者和使用者，也要通過倫理教育和職業倫理培訓，加強「軟」約束。

科學設置監管「紅線」也是當務之急。全國政協委員、中國科學院院士、南京大學黨委書記譚鐵牛在今年全國兩會期間就建議，需要明確人工智能在生成內容過程中侵犯他人的責任界定，明確應用紅線，並全面建立人工智能生成內容審核標準的評估驗證機制，確保人工智能內容生成技術安全、可靠、可控。




整理：香港文匯報記者 張帥

## 偏見信息氾濫或加劇國際衝突

**專家解讀**

中國傳媒大學政府與公共事務學院院長周亭稱，國際互聯網中充斥着海量未經過濾、包含偏見的信息，其中以種族歧視和性別歧視居多。虛假信息和偏見性信息如果通過生成式AI在世界範圍內大量生產和流動，甚至將會加劇分裂主義、種族偏見等國際衝突。

據英國《衛報》上周報道，其社內一位記者3月份收到一封電子郵件，一名研究員在進行研究時查找到了一篇文章，顯示是該記者在幾年前所寫，但在《衛報》網站中卻搜不到這篇文章，因為文風和主題都高度雷同，這位記者也不肯定到底是不是自己的文章。最終報社對所有的報道紀錄進行回溯，最終確定這篇文章根本不存在。

生成式AI不僅能冒充記者寫文章，其生成的圖片也難辨眞偽。日前，一幅美國前總統特朗普被全副武裝的紐約防暴警察按倒在地的圖片在社交媒體平台上氾濫，然而看似細節豐富的圖片卻與事實毫不相干，它們都出自人工智能驅動的圖像生成技術，其高度複雜的算法使得生成的圖像在視覺效果上極為逼眞。

生成式AI產品引發的爭議和擔憂不斷發酵，有關虛假信息氾濫問題受到普遍關注。中國傳媒大學政府與公共事務學院院長周亭稱，國際互聯網中充斥着海量未經過濾、包含偏見的信息，其中以種族歧視和性別歧視居多。例如，英國媒體Insider報道稱「ChatGPT曾告訴用戶可以折磨某些少數民族人士」。生成式AI這樣的回答基於對互聯網信息和受眾反饋的學習，不但影響認知，而且強化用戶的偏見。

周亭還引用美國一家研究機構的發現指出，如果對ChatGPT提出充斥陰謀論和誤導性敘述的問題，它能在幾秒鐘內改編信息，產生大量令人信服卻無明確信源的內容，如果此類信息通過生成式AI在世界範圍內大量生產和流動，甚至將會加劇分裂主義、種族偏見等國際衝突。

◆香港文匯報記者 張帥 北京報道

## 美國考慮針對AI工具進行規管

**香港文匯報訊** 據《華爾街日報》報道，美國拜登政府已經開始研究是否需要限制ChatGPT等人工智能（AI）工具，因外界日益憂慮AI會被用作歧視或傳播有害資訊。

據報，美國商務部日前就其所謂的問責措施正式公開徵求意見，為潛在監管踏出第一步，諮詢為期60日。當中包括具有潛在風險的新AI模型在發布前，是否要通過認證核准程序。

美國商務部轄下的國家電訊和資訊管理局（NTIA）負責進行是次意見徵詢。NTIA負責人Alan Davidson表示，當局認為需要為AI技術設置護欄，確保相關各方以負責任的方式使用它們。

Davidson表示，收集到的意見將用來幫助制定提交給美國決策者的應對AI建議，又強調NTIA的法定任務是向總統提供科技政策建議，並非編寫或執行法規。

此前，美國總統拜登在白宮與一個科學家顧問委員會討論了AI話題。被問及該技術是否危險時，拜登稱還有待觀察，有可能是危險的。

美國金融部門的監管機構已在調查貸款機構可能利用AI來審批貸款的做法，目的是防止對少數族裔的歧視。

美國司法部的反壟斷部門則表示，正在監測AI領域的競爭；美國聯邦貿易委員會（FTC）則告誡各公司，如果對AI產品作出虛假或未經證實的聲明，企業可能會面臨法律後果。

在日前發布的公共意見文件中，NTIA諮詢民眾是否應當增加措施如「質量保證認證」，以確保公眾對AI系統的信任。該文件亦諮詢民眾應否制定對AI適用的新法律或法規，但沒有詳細說明AI的潛在危害，也沒有表示支持任何具體的保障措施。