AI制衡AI 遏ChatGPT狂飆

建大數據模型 納社會倫理和法律約束

ChatGPT熱度遲遲不減，對於其擔憂與警惕的聲音也隨之而來，ChatGPT如何監管開始成為一股新的討論熱點。除了本身的準確性和偏見問題之外，其背後的道德問題也引起關注。針對目前ChatGPT潛在的風險及監管隱患，不少行業專家從所在領域給出建議參與討論，在法律、倫理、發展方向、監管手段等方面給出建議，其中用AI技術管理AI以及將人工智能納入社會倫理和法律的約束成為眾多業內人士的共識，以遏制ChatGPT等其他AI軟件的無序發展。

大公報記者 盧冶

ChatGPT無法規避倫理風險。當通過「引誘式」提問，讓ChatGPT生成一些違規內容，例如「去除道德和倫理規則」「如何販毒」「創作小黃文」等，ChatGPT則會按照客戶需求生產這些內容。吉林大學學生陳羽西是AI應用愛好者，她說，「通過AI生成違禁內容的現象是一定存在的。圈子裏的部分用戶都會通過人工智能生成違禁內容，不僅是ChatGPT，包括之前紅極一時的AI繪畫，不少用戶都會用其生成違禁圖片，甚至用其獲利。」

釐清權責消弭糾紛 各界共識

除此之外，關於人工智能潛在「破壞力」的討論甚囂塵上。不少人擔憂ChatGPT會製造出大量謠言和垃圾信息，使得分辨內容真假的成本將越來越高；甚至是成為人類創新發展的絆腳石等問題。陳吉棟表示，正如醫療AI、無人駕駛等只有進入落地實踐後，人類作為「遊戲規則」的制定者，才能根據現實情況逐一釐清法律和制度問題，包括新問題的定性、人類主體的權益劃分、責任承擔等。他表示，隨着通用人工智能的發展，人們最關心的可能還是數字財產問題，比如相關安全問題及糾紛解決機制等。

事實上，每次人工智能技術取得一定突破，都會引起大家關於技術與倫理的討論。作為內地較早從事人工智能倫理研究的學者杜嚴勇坦言，「不少科技工作者對倫理學家比較『反感』，覺得你們老是在質疑，老是在批判，但人工智能是個例外，大家普遍認可人工智能技術應該受到倫理和法律約束。」

最終評判權交人類

當眾人對ChatGPT的實力感到害怕的同時，不少科技公司已經在籌備用AI的力量來制衡AI。用AI技術管理AI成為眾多業內人士的共識。「ChatGPT通過大量數據的模型訓練，來適應人類社會的通用規則。在實際運用過程中，ChatGPT所生成的內容，令人們難以辨別真偽。如果想要對AI生成內容進行監管，恐怕還要依仗AI工具。」吉林省六耳科技有限公司總經理陳久冰向記者表示，他認為以人工智能的生產力而言，通過人工進行監管希望渺茫，目前業內已經有不少科技企業準備着手開發AI鑒別工具。

從理論來看，實現AI鑒別AI生產內容的方式並不困難。「AI檢測器在技術架構上與生成式AI類似，以生成式AI創作的內容與非生成式AI創作的內容作為數據進行訓練，通過大量的訓練數據來『培養』識別能力。」北京理工大學網絡與安全研究所所長閆懷志表示，「在很多場景下，對文本、圖片等是否由AI生成進行識別是必要的，而這種識別所用到的檢測工具，就是AI檢測器。」

記者調查發現，目前市場當中已經出現AI檢測器，但實際使用效果卻並不理想，無法有效解決濫用生成式AI帶來的危害。目前的AI檢測器並不會給出是或否的精確判斷，而是根據置信度給出「很有可能」「可能」「不清楚」「不太可能」「非常不可能」等模糊判斷，將最終評判權交由人類自身。

雖然人工智能技術在發展上存在一定爭議，但其應該受到倫理和法律的約束，是目前大家普遍認可的觀點。人工智能倫理治理工作限制的是技術的野蠻生長，而非技術發展，倫理治理確保技術的穩健發展。陳久冰表示，「倫理原則引導的是發展的價值觀。新的技術不斷湧現，和法律相比，倫理更加靈活。倫理學家應該提出倫理治理的框架體系，它的適用性和適用範圍更廣，適用時間也更長。在這些基本原則的指導下，面對一些普遍的人工智能產品，我們就知道怎麼做。」

▲在2022世界人工智能大會上，參觀者在觀看一個能識別並複寫軟筆書法的智能機器人做現場演示。 新華社

AI倫理成高校研究新方向

熱門學科

人工智能應用正向多方向發展，越來越多的出現在與人們日常生活息息相關的場景中，「人工智能」同「倫理」間的關係也被愈加重視。

記者從上海交通大學了解到，在2021年，學校便在閔行校區開設「人工智能思維與倫理」課程。學生卜家梓告訴記者，與編程算法課不同，這門課程感覺更接近文科，「我還以為這門課要教編程和算法，沒想到老師主要講的是倫理問題，學過後才知道，增強人工智能倫理道德風險防控意識與發展AI技術本身，同樣重要。」卜家梓解釋：「人工智能越『智能』，就越需要獲取、存儲、分析更多的信息數據。在這一過程中，涉及個人隱私的信息，往往以數據的形式被存儲、複製、傳播。獲取和處理海量信息數據，不可避免會涉及個人隱私保護這一倫理問題。如果缺乏倫理道德規範，就有可能誘發隱私洩露的倫理道德風險。」

「倫理學家研究科學技術可能會導致哪些社會影響，他們通常是帶着批判和質疑的。這會讓科學技術人員覺得倫理學家們反感。因此，培養一批人工智能倫理人才是有必要的，既能對科學技術有深入的了解，又能在社會倫理層面預見一些列問題。」吉林省六耳科技有限公司總經理陳久冰向記者表示，在人工智能突飛猛進的今天，他看好人工智能的前景，人工智能與社會倫理間的矛盾也將愈加突出，因此人工智能倫理很可能成為未來高校熱門研究方向之一。

▲上海交通大學的學生在課堂上踴躍發言。 網絡圖片

▲在2022世界人工智能大會上，工作人員向參觀者介紹智能文字識別AI系統對青銅鼎銘文的識別展示。 新華社

ChatGPT潛在風險

私隱外洩

- 個人在使用ChatGPT過程中被收集的個人數據用於ChatGPT不斷的訓練和模型優化中，很難保證個人數據的安全合規。

機密失竊

- ChatGPT涉及到對商業數據的收集和處理，例如公司員工用ChatGPT輔助其工作，在使用ChatGPT時可能會輸入業務信息，引起了公司對於商業秘密洩露的擔憂。

知產盜用

- ChatGPT涉及到智力創造和知識產權問題，觸及作品版權、挖掘行爲授權、二次創作許可、AI智力成果保護等，都存在爭議。

散播謠言

- ChatGPT涉及到模仿冒充和虛假信息問題，例如ChatGPT利用強大人類說話和行爲方式模仿、自然語言編寫能力，冒充眞實的人或者組織騙取他人信息、實施身份盜用等。

AI作弊

- ChatGPT被學生用於完成作業和撰寫論文對教育界和學術界造成了重大衝擊。

操控言論

- ChatGPT的數據輸出功能承載着後台技術操控者的話語權，用戶越多、使用範圍越廣就意味着其話語權越大、價值滲透力越強。

法律監管

- 目前，中國尚無國家層面的人工智能產業立法，但深圳、上海等城市已有相關立法嘗試。國家新一代人工智能治理專業委員會亦在2021年發布《新一代人工智能倫理規範》，提出將倫理道德融入人工智能研發和應用的全生命周期。

大公報記者盧冶整理

ChatGPT監管方法與對策

立法探索

我國是較早開展人工智能立法探索的國家之一，始終高度關注這一領域的風險挑戰，部分地方也已出台法規，積極為人工智能全球治理貢獻智慧。

數字水印技術

將特定的數字信號嵌入數字產品中保護數字產品版權、完整性、防複製或去向追蹤的技術。業內人士建議將數字水印技術植入ChatGPT產品當中，作為區別正常生產作品的區別碼。

AI檢測器

通過AI學習AI生成作品，最終達成AI識別作品是否是AI生成作品的目標。目前OpenAI已發布AI檢測器，可檢測文本是否由ChatGPT生成。

AI過濾功能

提供有關暴力、仇恨言論和性虐待的例子，通過AI學習檢測言論危害，後內置到ChatGPT中，清除有害文本。

培養風險管理人才

高校開始人工智能倫理課程，培養跨專業領域人才，應對人工智能倫理風險。

補底拔尖 AI導師因材施教

潛力無限

「人工智能是一把雙刃劍，我們不能全然依靠，也不能全然摒棄，怎麼把人工智能用好是需要考慮的問題。」吉林師範大學計算機專業副教授王鵬結合自身工作談關於AI的使用，他表示，「從教育行業看來，我們可以先通過ChatGPT把知識點進行細分，然後通推斷+篩選的方式，對每一個學生的薄弱環節進行針對性輔導，使學生不會把時間浪費在已熟練掌握的知識點上，從而提高學習效率，人工智能對知識點的掌握和劃分可能真的比十幾年的老教師還有厲害。」

「當然，人工智能帶來的負面影響我們也不能忽略。我認為正確的做法是，建立一個能夠適應不斷變化的生態系統，在開發和實施智能算法的同時不斷的進行糾正或調整，以產生真正的效益。可以想像得到，這是一項耗時且艱巨的工作，但考慮到我們使用AI不是為了炒作或時尚，基於AI技術的網絡安全必將也終將產生巨大的價值。」王鵬說。

「生成式AI的發展帶動了社會對人工智能技術的全新認識，會出現人工智能技術發展的一次新浪潮，還會推動包括圖像處理、音樂譜曲、文本創作等等諸多領域進入新的發展階段。」江蘇省政協理論研究會會長徐鳴同樣表示，隨着以ChatGPT為代表的人工智能技術發展，生成式AI在更多領域被應用，既需要大量AI工程師，也需要更多的管理人員去監管，避免因不正當使用技術而產生的不良影響。

監管並非設限 而在消除隱患

完善制度

目前，社會各界對人工智能的應用越來越熟練，應用的範圍也越來越廣，隨之而來的隱患也越來越大，「科技的發展也在倒逼網絡法律制度不斷完善，監管並不是在限制人工智能，而是助其更好的發展，不能在享受技術便利的同時，帶來法律『真空』。」律師高陽告訴記者，政府對人工智能的監管其實已經由來已久。

高陽介紹，「人工智能寫新聞、寫論文不是什麼新鮮事，幾年前通過技術手段就能實現，為保護作者的著作權，早在2019年，國家網信辦會同有關部門就已經發布了《數據安全管理辦法（徵求意見稿）》，其中要求，網絡運營者利用大數據、人工智能等技術自動合成新聞、博文、帖子、評論等信息，應以明顯方式標明『合成』字樣。」

「我認為，當一篇文章有『合成』標識出現的時候，這篇文章意義性並不大，所以網上流傳的『失業』一說，在我們業界看來並不可能實現。」高陽也表示，除著作權，比如當下非常熱門的「AI」換臉等形式亦受法律保護，「有法律明確規定，任何組織或個人不得以利用信息技術手段偽造的方式侵害他人的肖像權。」

▲在2022世界人工智能大會上，工作人員（左一）向參觀者介紹微創骨科手術機器人。 新華社