通用人工智能會預設一種“公理化”的普遍性嗎？*

吳　靜

[提　要]　通用人工智能的技術基底預設了一種生成知識普遍性的機制。通過可通約的大語言模型，將人類基於自然語言處理所形成的知識生產範式還原為對經驗數據的推崇，由此形成一種建立在數字模型和數據基礎之上的知識整體性結構。然而，大語言模型所暗含的對語言和經驗的可通約化理解，不僅會導致人類盲目信賴技術，將認知方式和技術框架相等同；而且也會使知識生產呈現出封閉化和中心化等趨勢。這種知識的生產特徵實際上代表了被技術通用性所預設的公理化知識體系，其本質是資本主義生產邏輯全球布展的體現。對通用人工智能的普遍性知識生成過程進行切實反思，必須深入到對隱匿在技術背後的公理化構架的批判性分析之中。

[關鍵詞]　通用人工智能　ChatGPT　通用性　可解釋性陷阱　公理化

[中圖分類號]　C39　[文獻標識碼]　A　[文章編號]　0874 - 1824 (2023) 04 - 0123 - 10

隨著以 ChatGPT 為代表的生成式人工智能（AI Generated Content，AIGC）朝向通用人工智能（Artificial General Intelligence，AGI）的強勢轉向，技術的通用性和作為人類共識的知識構型之間的關係成為反思技術問題的切入點。一方面，在深度學習的算法預演和龐大數據庫預訓練的基礎上，通用人工智能所形成的知識生產模式，以開放性、高算力和強互動性的特徵確保了知識傳播和獲取的普遍有效性和可讀性，為實現知識的公共性生產提供了可能；另一方面，由 ChatGPT 等 AGI 技術所反映出的知識構架實質上代表了一種中心化的“數字普遍理性”。這種知識結構儘管可以利用及時更新的數據庫而確保新文本內容的生成，但從生成機制上來看，它仍需要預設某種明確的知識普遍性構架（樣本模型、數據參數）作為形式保障。而一旦被普遍形式所中介，知識朝向未知領域的無限可能性便煙消雲散，由此成為被隱匿在 AI 技術背後的資本邏輯所操縱和預設的對象。

一、大語言模型與可通約性

在希臘神話中有兩個形象都與對話和語言有著密切的關係：一個是赫爾墨斯，眾神的使者，負

* 本文係國家社科基金項目“德勒茲資本批判視域下的西方平台資本主義研究”（項目號：20BZX011）的階段性成果。

責傳達宙斯的旨意;一個是斯芬克斯,用繆斯所傳授的關於"人"的謎語向過路的眾人提問。前者是權威信息的單向傳達,後者則是以提問的方式等待答案。有趣的是,這兩者都不是真正的對話,因為接受一方並沒有參與到整個意義生成的過程與結構中來,雙方的輸出始終處於不對等的地位。然而,這兩處互動都將信息的表達以最接近自然性的方式——人類語言(這正是人類文化形成的基礎)——呈現出來,在傳遞者與接受者之間形成了無障礙的閉環。在這種信息傳達的過程中,一種可通約性的預設隱而未現:無論作為"對話"另一方的是奧林匹斯的眾神,還是不知來自何方的行人,他們對赫爾墨斯或斯芬克斯所言說的意義以及傳達意義的語詞本身都並未提出異議,信息理解的門檻幾近於無。當然,與此相關的兩個主人公形象本身也頗具相當程度的文化投射意味:赫爾墨斯擅長雄辯(語言的藝術),而斯芬克斯在希臘神話中則象徵著智慧與知識(語言所表達的內容)。於是,實現可通約性的兩個制約性方面都以前提的方式得到了解決:"腦"和"口"——理解和交流,它們分別對應著意義和語言(包括其所表達的情感),或者說所指和能指體系的集合。當然,對於通用人工智能技術的願景來說,這種可通約性只是相對容易實現的第一層次目標,它還需要在知識表現、推理、學習、交互甚至執行等多個方面有突破性進展。儘管目前新的人工智能工具的通用性還遠未達到這些里程碑的要求,但這並沒有妨礙很多業內人士將作為一種自然語言文本生成系統的 ChatGPT 的問世視作通用人工智能的奇點。雖然 GPT 系列的代碼生成過程基於訓練文本,但由於其底層技術應用了自然語言處理系統和大語言模型,GPT 系列可以以對話者理解的自然語言方式為行動提供指導性意見、問題預測甚至解決方案,從而展現出激發人類創造力的巨大潛力。這也就是為什麼尤瓦爾·赫拉利在 4 月 29 日的 Frontiers 論壇上的演講中特別強調了如果將通用人工智能的能力實現歸結到一個更具有決定性的基礎,那麼這一基礎就是操縱和生成語言的能力,無論是用圖像、聲音還是文字。赫拉利將人工智能對自然語言的掌握視作人工智能革命的一個重要方面,這誠然一方面是因為語言本身關乎人與人工智能連接的交互界面和友好程度,但另一方面則關乎人工智能的功能"通用性"的實現與語言所構築的關係親密度和社會邊界。更準確地說,在最友好的關於人—AI 交互界面的想象中,功能的"通用性"(即多目的性)必須與語言及所表達的意義體系的可通約性形成適配。當然,赫拉利的這個宣稱本身也以隱喻的方式回應了聖經《創世紀》的第一章:創世的初始行為即是以語言進行的——上帝說"要有光",於是有了光。

其實人工智能所依賴的自然語言處理(Natural Language Processing,NLP)系統與人類對自然語言的研究和處理方式相去甚遠,其目的是使計算機系統以人類語言為中介,有效地實現和用户之間的通信。它對語言的掌握更多地是沿著數學、統計學和語言學的框架進行,而不關注單個語詞的含義和語言結構的意義生成。這產生了一種方法論上的悖論:一方面,知識本身被所有浮於表面的經驗數據(包括網絡文本的經驗生產)建構,由於 AIGC 和其底層技術的大語言模型都並不具有真正的"思考"和"理解"的能力,它只能通過深度學習的思維鏈條和語言概率邏輯實現文本生成。因此,被生成的文本更多地是數學邏輯的結果,而非綜合的認知能力。這和人類認識所希望達到的"認識你自己"和"理解世界"的目的相去甚遠。1930 年代以來認知計算主義與反計算主義曾經有過激烈交鋒,爭論的焦點在於對世界的理解是否可以完全以計算方法來解決。不過,當時計算主義所面臨的很多詰問在今天大數據和大模型的技術應用得以實現的情況下已經得到了部分的解決。以 GPT 系列為例,當數據類型和數量以及模型參數呈現指數級增長後,通過深度機器學習所獲得的結果(如 GPT-4)就已經在一定程度上顯示出和人類思維方式、甚至情緒方式的相似性。樂觀的計算主義者相信,從任何可證偽的意義上來說,知識甚至語言的統計學則相當於理解。而大語言模

型所實現的正是這種意義上的知識獲得過程和理解過程,其數據覆蓋的全面性保證了這一過程的可靠。這種對計算主義認知構架的篤信使得 AIGC 的生成內容越來越容易獲得人們的信任,而作為其根基的數據和模型問題則被忽略。

大語言模型實際上是利用海量無標註數據訓練而成的深度神經網絡模型。ChatGPT 正是在大語言模型基礎上運用 Transformer 架構的自回歸模型,它通過預訓練和微調兩個階段來進行自然語言的學習和處理。在預訓練階段,GPT 系列使用規模極其可觀的無標註文本數據作為初始輸入,通過掩碼預測任務來學習文本中單詞之間的語義和語法關係。進入微調階段,它則根據不同的下游任務(如提問、編碼或文本生成等)進行少量樣本的學習,從而適應特定領域、風格和任務。這種生成式人工智能的學習能力取決於預訓練參數的規模,參數的規模決定了生成文本的質量:它不僅表現在生成文本的信息量和準確度上,也表現在語言的"親人性"上。這也是生成式人工智能與單純的搜索引擎之間的關鍵差別。從公布出來的數據得知,GPT-2 大約有 15 億個參數,而 GPT-3 最大的模型有 1,750 億個參數,相較前者足足上升了兩個數量級。根據媒體報道(但還未被證實)的消息,GPT-4 的參數可能將達到 100 萬億的規模。這也就要求必須將更多的數據投入到模型訓練之中。從認知計算主義的意義上說,大語言模型在邏輯上擔當了一種知識整體(儘管它自身也在動態變化)的功能,打破了過去人類知識傳承和學習的分散性,從而為認知的中心化提供了前個體層面的可能。和之前很多同類產品不同的是,應用大語言模型的 GPT 系列不需要針對每個對話場景設計特定的規則或模板,而是通過給出提示、上下文信息或目標指引的方式,就能生成流暢、自然、有趣的回覆。在 ChatGPT 宣布可以集成第三方插件、實時聯網之前,它的訓練數據集僅截至 2021 年 9 月。而在這一封印解除之後,GPT 系列不但可以使用即時的網絡資源,還可以和網站互動。這也就意味著數據庫的不斷更新,更多、更即時的數據會加入到第一階段的預訓練中。從理論上來講,這有利於訓練出更好的模型從而吸引更多用户的加入和使用,以至於產生更多用户的數據和模型參數並用於進一步的訓練,形成良性循環。這就是數據"飛輪效應"。在這個過程中,數據和模型的增長形成了相互促進的關係,並且隨著時間加速效應越發顯著。

此種規模的數據庫是之前任何預設目的的數據採集都難以實現的。在這個基數量級之上借用數學統計方法所實現的可通約性遠超過絕大多數模型。這種可通約性被數據分析和應用數學視為保真度的保證。通用人工智能的保真度越高,人和它之間的"親密關係"的強度就越高。這也就意味著,大語言模型(Large Language Model,LLM)的海量數據庫所推進的自然語言處理系統的可通約性和親人性(赫拉利將之形容為人與人工智能的"親密關係"建立)加強了人工智能在可解釋性方面獲得的權威感。在數字化技術與 AI 技術的應用和發展中,可解釋性一直是連接經驗與其數字化表示之間的橋樑。只有當可解釋性成立的時候,數字化表面才能獲得其存在的合法性。它一度被認為是數字技術可信度的正面展現。但佐治亞理工學院的研究團隊在 2021 年的一項實驗中提出了"可解釋性陷阱(EPs,即 Explanibility Pitfalls)"的概念,旨在尋找數字技術可解釋性的負面影響以及對可解釋性信任程度的影響因素。[①]這項實驗發現,具有不同教育程度和背景(對數字技術和人工智能技術的了解程度不同)的用户都對數字設備所輸出的結果表現出了盲目信任的態度,並進而產生出依賴性(啟發式信任)。儘管研究也表明導致這種盲目信任的原因並不完全相同(語言只是其中的一個因素),但人與人工智能的關係建構明顯具有權力分布上的不對稱性。這種過度的信任使得算法和模型的問題不但在技術黑箱的層面上、更在本質的層面上難以得到有效的審視和監督,極有可能陷入"可解釋性陷阱"(EPs)。從另一方面來看,大模型暗含的語言和經驗的可

通約性也會加重“可解釋性陷阱”的存在。原因在於,經驗的可通約性的背後實際上是以全球化為基礎的現代性社會生產方式的布展,它體現的正是特定文化的中心化,具有地方區域性或非“主流”色彩的信息在可解釋性問題上明顯不佔優勢。通用人工智能的“可解釋性陷阱”不但使得認知方式和技術框架更加趨向一致,源於自然主義的科學主義態度更加深入人心;同時自然語言貼合度的提高使得整個人機互動過程體現為自然的社交過程,加深了人機的親密度以及由此而產生的人對技術的信任感,使得人工智能生成的文本(語言)成為具有更高可信度的“私語”或“話語”。當然,“私語”和“話語”的可信度所依賴的基礎並不相同,前者是互動雙方之間的親密關係,後者則是發布方的權威性。生成式人工智能同時實現了這兩方面的突破。

有研究表明,基於機器深度學習、以社交方式出現的人工智能服務體(Artificial intelligence service agents,AISA,生成式人工智能和通用人工智能都屬於這種應用)對於用户所表現出的共情反映有助於加強用戶對其依戀和信任。[②]不過,和現實人際交往中的親密關係的排他性和私密性不同,生成式人工智能生成的“私語”和“話語”作為網絡文本資源都會成為公共知識空間和公共領域的組成部分。這種公共空間的連接很容易讓人聯想起馬克思在《1857—1858 年經濟學手稿》中首次使用的“一般智力(general intellect)”範疇,這一範疇是指工業資本主義時代的大機器生產模型中,工人通過使用機器所形成的人—機協作關係,其在本質上是資本主義生產方式的內在矛盾的產物,被作為現實社會過程的“直接器官”。而意大利自治主義學派的維爾諾和奈格里等人則在“非物質勞動”的維度上重新定義了該範疇,使其超出了馬克思語境中對固定資本的條件性指認,而轉變為勞動主體所普遍共有的潛能(如認知、語言、情感及反思等),並以此為基礎改寫了福柯和阿甘本的生命政治理論,主張諸眾應通過發揮一般智力形成協作性的生命政治生產來推翻資本的霸權統治。維爾諾甚至特別強調了語言合作在這種協作關係的實現中的特殊地位,它是人類智力被導入機器體系的途徑和方式,因為“正因為有這樣互相協調一致的部分,智力才沒被導入機器系統,而是以人類勞動的直接活動,以其語言合作來體現。”[③]在這裡,語言的突出地位明顯地強調了不同於工業資本主義時代的大機器使用的人機互動界面。不過,在最新出版的英文著作《世界的觀念:公共智力與使用生命》(*The Idea of World: Public Intellect and Use of Life*)中,維爾諾進一步將“一般智力”範疇改寫成了“公共智力”,旨在強調它對於“公共範圍(public sphere)”形成的作用。從一定程度上而言,AIGC 的文本生成是人工智能時代這種“公共範圍”的具象化,它同時也顯示出語言系統的可通約性和“普遍理性”及其表達形式(對話)對於知識生成的影響。

二、大模型的危機:“遞歸的詛咒”與可解釋性陷阱

史蒂芬·霍金曾提出決定人們對於現實認知的,並不是所謂的“真實”,而是關於“現實”的模型的觀點。這裡的“模型”並不是獨立於現實之外的先驗性實體,而是人的認知系統。經驗“事實”只有經過認知系統的組合和構型,才能被人認識。從這個意義上說,沒有不依賴於模型存在的事實。無獨有偶,吳冠軍在他的新作《從元宇宙到量子現實》中更是直言不諱:“我們所體驗的‘現實’,只是我們這個宇宙賴以構建自身的那組數學結構所開啟的‘關係性現實’。”[④]它構成了阿爾都塞在談論症候閱讀時提出的“柵欄”概念——決定了讀者視域的問題式框架。這些理論,儘管使用的概念和論述方式不盡相同,但其實從偵探小說寫作手法的角度去看就不難理解了:作者暗地裡鋪陳和透露出的事實“碎片”,只有在被納入了所謂“真相”的拼圖中時才顯示出自己的意義。而在那之前,關於所謂“真相”的不同敘事會以迥異的方式將它們連綴在一起。這其實就是霍金的“模

型”或阿爾都塞的“柵欄”。齊澤克則刺穿了“模型”或“柵欄”的實在論性質和這些範疇本身所帶有的強烈結構主義色彩,進一步地提出了這種決定性的模型或結構在本質上更依賴於其被表徵的方式。他將“符號之序(symbolic order)”指認為是從直觀感覺到先驗統覺之間被隱匿的關鍵性中介。在他看來,所謂的“現實”,其實只是被符號表徵建構起來的連續性景觀。“符號性的向度就是拉康所說的‘大他者’,那個將我們關於現實的經驗予以結構化的無形的秩序,關於各種規則與意義的複雜性網絡,它使得我們看見我們所看見的——按照我們看見它的方式(它同樣使我們看不見——按照我們看不見它的方式)。”[⑤]這也就意味著,不但大模型所生成的文本輸出,甚至連輸出內容被表達的語言和方式都在決定人們對“現實”的理解,從某種意義上說,也就決定了“現實”本身。然而,需要注意的是,這裡問題的關鍵並不在於表達的內容與表達形式的自主性之間的關係,而在於前語言層面所形成的、決定了“說出的和未說出的”、“看見的和看不見的”思維體系。

由此可以理解,當問題由對象本身轉而回到對象的表徵維度時,尤瓦爾·赫拉利對於語言在人工智能與人類文明系統之間所扮演的角色的擔憂其實並沒有那麼聳人聽聞。它不過是提出了語言作為符號性中介對人類認知和文化表達的影響。而這一點,維特根斯坦早已做過論證,他的《邏輯哲學論》就試圖對語言和現實之間的關係作出辨析,並為思維的表達劃定界限。其中的著名論斷“語言的邊界就是思想的邊界”,並不是說在語言的符號表達之外不存在思想,而是意味著,語言的符號系統的解釋能力給定了思想的基礎,也奠定了人們思考和言談的方式。“思想在命題中得到了一種可由感官感知到的表達。我們用命題中的可由感官感知到的記號(聲音的或書寫的記號等等)作為可能情況的投影。投影的方法就是思考命題的意義。”[⑥]記號體系所體現出的“被投影者的可能性”正是語言表達的邊界。只有在這個邊界之內,對於命題的表達才是可能的。儘管維特根斯坦的這一表達是針對命題的真值判斷而言,是以一種邏輯實證主義的方式所描述的事實與語言命題之間的邏輯關係。他所強調的邊界也是邏輯表達的邊界。但這一論述同樣可以被延伸到形而上學意義上語言對思想的構型意義:語言所給予的概念、概念之間的關係、語法形成的判斷、詞彙的差異及價值都為思想之所以成為可能提供了路徑。這當然並不意味著語言是某種先於事實或思維的實體性預設,而是從生成條件的維度上考察語言所表達的信息與“事實”之間的關係,它挑戰了語言單純作為思想/知識表達工具的觀點。

在這樣的前提之下,從算法的底層邏輯出發對大模型所提供的知識圖景進行審視就是一件十分必要的事情。它可能對未來的人類思維和認知都發生重要且不可預測的影響,甚至有可能反饋在每一個單獨的判斷中。對此研究者需要格外警惕的不僅僅是保真性的問題,更在於這種知識生產方式所生產出來的“普遍性”知識是否隔絕了多元化的可能。有研究表明,一旦某種模型或認知方式與主體有限的觀測(上文已經論證,這本身就有敘事編織的成分)相符,或是有效地維持了我們的生存,觀測者就會對其產生強烈的信任與依賴,以至於排斥其他模型和認知方式,從而喪失對所獲得信息的批判性思考能力。

來自牛津大學、劍橋大學、倫敦帝國學院、多倫多大學等機構的研究人員通過實驗和測試,發現了基於大語言模型(LLM)的生成式人工智能在使用自身生成文本訓練模型時,出現了“模型崩潰”的現象。[⑦]這是機器學習的一種退化式生成。原因在於,生成式人工智能創造文本的能力遠遠高於人類,新生成的文本在數量上會呈現出激增的趨勢。同時,由於這些生成文本缺乏自我標註,會重新作為數據來源進入新的生成過程。這種不加審視的生成數據最終會污染下一代模型的訓練集。而使用被污染數據進行訓練,則會導致模型誤解現實,從而使知識的對象在生成式人工智能的技術

條件下被凝固。其中,在特定條件下的晚期模型崩潰中,模型將原始數據分布的不同模式相互引用、不斷“反芻”,最終會使新生成的數據分布收縮到與原始模型態勢相差甚遠的情況,呈現出較小的方差。按照這一趨勢,“反芻”的次數越多,方差就會越小。研究者將這種情況稱之為“遞歸的詛咒(the curse of recursion)”。這種方差上的縮小意味著新生成結果對異質性的表達減弱,知識越發呈現出中心化的趨勢。它同時也說明,人類傳統知識生產的自主性特徵被顛覆,知識生產不但沒有因為數據的擴展更多元化,反而樹立起新的封閉性邊界。這種封閉邊界和經驗時代知識生產所展現出來的局部性和區域性不同,它並不是對知識進行有條件的表述,而是對知識的內容進行不斷收縮的校正,最終和真實相去甚遠。從某種意義上說,這是一個消除“異議”的過程:由於大模型的底層算法主要與統計學上的概率相關,概率越小、“越不可能”的情況在重複投餵訓練的過程中會持續被忽略。換言之,當大模型自我生成了某種對現實的表徵後,它通過“自引”的方式不斷將這種表徵向內投射,從而在後續的輸出中不斷強化這種由其自身建構起來的投射。原始數據的離散度越高,“反芻”式數據訓練形成的方差距離原始方差或所謂的“現實”的距離就越遠,知識中心化的程度就越高。其結果就越不能形成有效表徵。這種事件“顆粒度”(即數據離散程度)的衰減甚至消失,仿佛是通過標準差的冪次方的方式排除了偶然性,創造了一個決定一般經驗所有條件的先驗領域。這種概率上的綜合和康德哲學意義上的人類先天綜合判斷不同,它既隔絕了和經驗之間的關係,又不具有普遍必然性,而是一種特定表徵方式(如 AIGC)所造就的知識的擬像。證偽這種知識擬像的難度隨主題而異,所以探尋對其進行檢驗方式的意義並不大,而應轉過頭來認真審視其生成機制:以大模型為底層技術的人工智能模型生成的數據在越來越具有普遍性的情況下,是否正在製造某種關於世界的均一化或公理化知識體系?這種知識的公理化又在何種程度上重構甚至改變著人類對於現實豐富性的理解?

也正因為如此,有學者對大模型技術的過度使用提出了根本性的質疑。他們認為,依賴於海量數據的大模型並不是解決所有人工智能技術的萬應良藥。因為大模型所依賴的自回歸算法不但需要耗費巨大的算力資源和長期的訓練時間,並且有可能產生欠擬合(訓練不足)或過度擬合(訓練過度)等問題。前面談到的模型崩潰也可被視作過度擬合的一種形式。過度擬合的可能性不僅取決於參數和數據的數量,還取決於模型結構與數據形狀(即數據分布的離散度)的差異大小。AIGC 的生成文本一旦進入自己的訓練數據,過度擬合的收縮必然難以避免。對這個問題的補救往往需要訴諸於更大更全更新的數據庫進行多次驗證,這種循環顯然是非良性的。事實上,在很多任務的實現上,數據的數量並不是保證模型適配性的唯一因素。相反,數據的質量、可靠性以及模型適用性的標準同樣重要。在某些情況下,小數據集可能更加準確和可靠,因為它們更容易進行有效的數據清洗和篩選。弱算力的系統也可以通過使用高效的算法和優化技術來提高性能,如可以使用並行計算、分布式計算和硬件加速等技術來提高系統的效率和性能。而且,和大模型技術後期的“遞歸詛咒”相反,小數據學習進路在後期隨著數據集的增加和模型的優化,可以取得更好的效果。這是因為小數據學習更側重於深入理解數據和模型,通過精細調整和優化模型架構、特徵工程等方面,取得更好的效果。與此同時,還要關注人機交互系統中用户的獨特性與價值觀,讓機器能夠理解和適應這些特異的性質和價值觀所包含的文化內涵,從而使人工智能的輸出不僅僅成為單向的“獨白”甚或“神諭”般的規訓,而是形成有真實意義的信息回路。當然,這也是通用人工智能發展所面臨的巨大挑戰,因為“通用性”本身所預設的普遍性與對象的獨特性之間存在著明顯的張力。如何在兩者之間建立起有效平衡,既使得“通用性”本身不至於成為人工智能知識圖景的權力

話語,又使得特殊性的存在不至於干擾基礎結構從而影響“通用性”,是通用人工智能在解決功能和目的問題之外格外需要關注的問題。

這個問題同時又關係到人工智能的可解釋性與人機信任建立的基礎。良性的人機交互基礎是可解釋性、透明性以及由此產生的信任。人工智能的可解釋性是指為人工智能的系統行為提供人類可以理解的理由,它是人類經驗理解人工智能的直接性紐帶(但在現實中並不一定總是正面和積極的),良好的可解釋性往往帶來更高的可信度。透明性則可以幫助人理解人工智能模型在其決策過程中所做出的選擇,包括做出決策的原因、方法以及決策的內容,它是對技術和算法黑箱進行解密和“祛魅”的努力。在此基礎上的人機信任是數智有機融合的推進。以大模型為基礎的生成式人工智能(一度被認為是通用人工智能的雛形和部分實現)在這些方面面臨著新的挑戰和危機。因為它不但造成了更高強度的“可解釋性陷阱”,而且在使用中難以被發現和矯正。

澳大利亞阿德萊德大學的研究團隊發現,用户在與以文字或語音激活的虛擬助理或對象的人工智能應用進行互動的時候,由於界面製造了類似社會互動的感覺,可以使用户體驗到一種社會交往所特有的參與和反應的即時感,在心理上則會對雙方的長期持久關係形成高情感反應。[⑧]生成式人工智能所引起的關於主體性、意識和情緒等話題的討論,其實可以看做這種高情感反應(包括困惑)的表現之一。這種情感期待和可解釋性陷阱一起,成為除算法黑箱之外,形成人機權力不對稱的原因之一。當然,這裡的“陷阱”一詞並非意指某種有意圖的欺騙或策略,而更多是比喻性地表示未被懷疑的或不容易識別的困難或危險。這些困難可能是由於缺乏信息、理解或監督造成的。但如果不能有效識別和避免這些陷阱,用户的風險就會增加。

在上文提到的關於“可解釋性陷阱”的研究中,儘管測試對象的技術背景不盡相同,但他們都會因為以數字化形式出現的輸出結果而增加對人工智能的信任。[⑨]這種對於智能系統的過度信任和不正確的感知一旦借助通用性人工智能應用進入到社會層面,就會使得人類喪失特有的批判性思維,並進一步影響到應用下游,進而產生難以估量的負面效應。如果這一問題在未來的通用人工智能技術發展中不能得到妥善解決,再疊加上大模型的退化式“模型崩潰”,“通用性”極有可能在全球範圍内造成一種令人生疑的“普遍性”。荷蘭代爾夫特理工大學的研究團隊認為具有良好的可解釋性的人工智能在算法設計時需要注意,不同社會的規範和價值觀本身具有差異性,決策過程必須側重於使這些規範達成共識。[⑩]這種共識不是概率上的平均值或中位數(這會導致錯誤的“客觀性”),也不是分布上的方差縮減,而是不同可接受性之間的平衡與博弈。這意味著決策必須是基於數據和可能的多種演算法之間的複雜交互,目的在於得到更廣泛的受眾的理解。儘管即使這樣,也不能保證規避所有的偏見和爭議,但該提議畢竟刻畫出了理想的技術願景。值得注意的是,該研究還指出,由於算法和數據都不是靜止的,往往會從自己的前序生成和新數據中實現進一步的學習(如大語言模型的自回歸算法),其結果是,算法的上下文越動態,下游的輸出和可解釋性就越無法得到保證。這不但從側面驗證了所謂的“遞歸的詛咒”,並且說明大模型借助於類似本質主義的元邏輯,在形而上學的意義上,生產出所謂的“真實”,把握了對於世界的合法定義,並通過不斷地自我確認和再生產,將自身的邏輯常識化和永久化,成為新的“數字自然”。

三、被通用性所預設的公理化知識

人工智能技術的通用性和自然語言處理系統對於可讀性的要求從根本上預設了一種強傳播性的認知模式,其所隱含的對知識普遍化構架的青睞,和德勒兹對資本主義“公理化”的分析具有根

源上的一致性和同構性。AGI 對通用性的追求及其對於知識構架的預設是隨著資本全球化進程所產生出的普遍性生產範式的側寫,這種生產範式以消弭地區、種族、文化等任何不能被資本一體化運作框架所涵蓋的差異性因素為首要原則,是資本總體化邏輯的具體展現。德勒茲曾以“公理化”(axiomatization)來形容資本主義體系的運作機制和擴展趨勢,相較於前資本主義社會借助特定符號,將社會要素的流動限制在既定場域、憑藉明確且不可逾越的規則實現對社會進行整合的“編碼化”體系,資本的公理化不但具有更為嚴密和周全的特徵,且借助著強大的同化(資本化)邏輯將所到之處的一切元素裹挾進自己的洪流。它並未取消對社會要素的控制,相反卻通過更加遍在的、邊界更無跡可循的“強中心”體系強化了控制的深度和廣度:“只有一隻作為中心計算機的眼睛,它進行著全範圍的掃視”。[11]資本的公理化同樣體現在對知識的構型中。一方面,公理化的最大特徵在於中心的嚴密性和邊界的彌散性,這一中心僅服膺於以市場為核心的資本生產邏輯,並隨著現實條件的變化需要而靈活調整其邊界範圍,任何知識的形成只有在獲得資本調配的前提下才成為可能。另一方面,公理化構建起一套基於資本普遍生產的普適性準則,它“不提供任何評註或解釋,僅僅是一組有待實施的規則”[12]。在這樣一種消弭了任何意義闡釋和特異性關照的規則體系中,由於規則本身成為了不驗自明的“真理”式話語,規則背後的資本就得以建立起壟斷一切信息和知識來源的“能指的霸權”。

藍江認為人類用户和 ChatGPT 之間的交互,涉及的不僅僅是認識論問題,而是一個將對象進行虛構構型的本體論問題。只有在完成這種想象性構型的前提下,人機交互才能發生。[13]這一觀點其實過分強調了交互界面中主體一端的心理選擇機制,而忽略了數字技術本身的座駕功能。事實上,通用人工智能反映了數字時代知識的公理化趨勢。儘管深度學習的算法構架、基於大數據提取與分析的預訓練式語言生成模型,從表面上看似乎消解了前數字時代知識獲取的過高准入門檻和複雜學習成本,呈現出“人人皆可參與,事事皆可獲取”的知識降維式生產過程。但是這種知識生產的技術基底卻始終秘而不宣,難以被公眾所了解。受限於技術研發與運作本身的不透明性,人工智能的“通用性”之中很難不摻揉進複雜的社會和經濟因素,而一旦技術的通用性被經濟和社會因素及其背後的資本運作邏輯所干預,它的“通用性”中立外衣便很容易扭轉成資本“公理化”的普適翻版。一個顯而易見的例子是“算法黑箱”問題:通用人工智能的程序設計、數據庫的揀選無疑會受到算法編寫者的價值觀、喜惡偏好乃至意識形態的影響,當編寫者將帶有個人評判標準的算法程序呈現在公眾面前時,這種個人偏見便會成為普遍性知識生產的現實基礎。類似地,受限於技術本身的不開放特性,由通用性技術所生產出的知識結構同樣可能具有封閉性、獨斷性的特點,社會中本已存在的意識形態偏見、甚至是技術的無意識會被無限放大且難以被察覺。更為重要的是,知識生產本應以去中心化的、朝向無限可能性的空間為目標,但 AGI 恰通過強中心點——生成式算法參數,和不斷調整的技術邊界——隨時可被擴充的網絡數據庫,抑制了知識的開放性生成。在數字時代互聯網成為人們絕大部分乃至唯一的信息獲取和知識學習渠道的背景下,這種被“通用性”所建構的知識正是資本的公理化在知識領域的新型展現。

技術的通用性所呈現出的知識公理化趨勢,同樣可以從“數字的普遍理性”所造成知識經驗性/差異性維度的消解中得到理解。在人文主義長久以來的傳統中,由“言說”和“觀看”等差異化方式所形成的經驗性共識是構成知識結構的牢固基礎。德勒茲就曾以“先驗的經驗主義”來探究經驗形成的條件,並將促成經驗自由生成的前哲學和個體式的“内在性平面”(plane of immanence),視作對抗主體哲學體現為“樹狀思維”的森嚴等級式知識構型的理論出路。通用人工智能對普遍性

知識結構的無意識追求,卻在理論上近似於一種向主體哲學知識結構的形而上學式倒退,它預設了數字這一形成經驗性共識的根本前提,並通過互聯網中無處不在的算法推送、數據生產強化了這一認識。公眾囿於數字媒介對信息獲取形式的壟斷,只能被動接受和理解經算法揀選後的信息。除數字(技術)之外,知識別無其它來源。人工智能技術的通用性將數字的地位推至極致,造就一種不驗自明、無可置疑、睥睨一切社會存在的"數字的普遍理性",消解了數字以外共識性經驗的生成可能。這種觀念不僅形塑著人們的知識獲取,而且也通過作為知識先驗性呈現場域的方式主導著知識的生產。當德勒茲說"並非是工具界定了勞動,正相反,是工具預設了勞動"⑭的時候,在某種程度上正是數字時代知識普遍生產的寫照。一旦公眾對由通用性技術所制定的數字普遍化和絕對化準則趨之若鶩,那麼知識的生產很難不帶有迎合數據流量、數字景觀至上的特徵。不難發現,由這種"數字的普遍理性"所主導的知識生產模式實際上與資本的公理化生產別無二致,它將一種普遍化的觀念架構無差異地置入知識複雜的生成過程之中,抽離了知識的情境性、獨異性等具體經驗內涵,從而借助普遍形式最大限度地保障知識生產的共識性。況且,由人工智能的通用技術所產生出的知識結構不僅遠離具體經驗,而且也通過普遍化架構的方式改寫著經驗,將經驗與數據等量齊觀,從而在由數字所中介的知識共識性生產中消解了差異經驗所可能蘊含的創造性、開放性的維度。在當代資本主義社會中,這種知識共識性生產便體現為資本要素對社會公共性的"總體吸納",其中技術邏輯和資本邏輯相互交織,共同構成一種壓制所有在場主體的無形同一化力量。

從更深層面而言,通用人工智能技術所造就的通用性知識構型之內存在著一對實質性的矛盾——知識形式的開放性和知識內容的封閉性。通用人工智能通過大數據預訓練所形成的深度學習模式,要求盡一切可能打破既有知識結構的封閉布局,將知識獲取的邊界不斷延展,朝向聯網數據源和大模型參數的前景不斷邁進。這也正是德勒茲所指認的公理化對有著固定意義和符號內涵的"編碼化"(coding)體系進行"去轄域化"(deteritorialization)的過程。知識的生產在這一過程中呈現出一種開放性的去中心化路徑,其生產的對象、範圍、方式均處於流動開放的生成圖景之中。生成式人工智能實現數據源截至到 2021 年 9 月的 GPT-3 到能夠即時聯網並可接入任何插件程式的 GPT-4 的跨越,其本質正是這樣一種去轄域化的知識構型的實際體現。但是,公理化同時存在著"再轄域化"(reteritorialization)的發展態勢,它要求將被解域化的要素重新限制在總體化程度更為嚴密的公理規則之中,以無所不包的邊界將多餘要素重新吸納入資本生產的整體性框架之中。這一點也正映證了 AGI 的知識生產特徵。一方面,通用人工智能的強中心算法模型樹立起一套封閉的知識結構,遮蔽了任何無法被數字所表徵的知識出場可能。不僅知識生產的內容受到算法參數的限制,而且知識生產的主體也時刻接受著來自"技術大他者"的無形型塑。儘管 OpenAI 公司宣稱多元式的參與和監督機制可以確保人工智能技術的真正開放性,但這種機制本身是否是新自由主義的技術翻版?被數字技術所遮蔽乃至阻絕於公眾話語之外的群體是否具有參與可能?這些問題促使人們不得不去反思隱匿在 AI 技術背後的資本主義生產關係。另一方面,通用人工智能進行知識生成的數據庫可能會將知識內容限制在數據編輯者所偏好的封閉空間之內,造成一種群體無意識的知識割裂和零散化分布,從而將隨著資本生產關係矛盾而擴大的階級固化、社會撕裂等社會不合理現實無限擴大。當下"數字繭房"等問題層出不窮的背後,是人工智能技術憑藉形式通用性所造成的結構封閉性的知識發展趨勢。在這種趨勢下,知識的生成形式(邊界)越是開放,它的實質內容(中心)就越是封閉。這和德勒茲對資本公理化運動的分析簡直如出一轍:"資本主義撞擊到了它的邊界,但同時又移動著這個邊界,將它置於更遠處"。⑮通用性由此成為一種和資本主義生產邏輯同構的知識產生範式。

通用人工智能所造成的知識公理化發展態勢，從某種程度上說正是資本主義意識形態在技術領域的集中反映，這種意識形態“將新技術同人類知識的發展相聯繫……它試圖通過計算機腦力勞動建立起一種新的人類集體主體性，這種主體性意味著人們可以僅憑一種‘數字—字母’語言而工作，從而擺脫身體的重負與差異”⑯。不過，當 AI 通過算法程式將知識轉變為碎裂的數字再現之時，與其說存在著一種建立於共識性基礎之上的意識形態話語，毋寧說一種社會性的技術無意識已經瀰漫和滲透在每一個應用環節。這種技術無意識割裂了知識與對象、經驗與認識之間的多樣性連接可能，取而代之以由算法所建構的單一數字型知識生產邏輯。和德勒茲同時代的技術哲學家西蒙東曾以“個體化”來形容技術發展的可能前景，這一理論要點在於強調技術物體內部各組成環節之間的協調互動、以及與外部情境相互影響且密不可分的一體化過程。然而就知識的生成而言，通用人工智能所依憑的算法“通用性”卻阻隔了技術與外在世界相互連接的真正可能，它無法感受人類經由漫長歷史文化沉澱所形成的知識構建過程，將數字確定為知識獲取的唯一且確定的來源，消解了對知識背後所蘊含的人類存在意義的追問。要刺穿這一困境，要求我們一方面以更加審慎的態度來對待通用人工智能的發展，通過持續性的反思和預訓練擴展，使 AI 技術超越“公理化知識”的窠穴，切實地為人類的未來造福；另一方面，切實地將豐富性和開源的原則引入到從技術設計到應用的環節，保持對單一技術、單一模型、單一數據來源的過分依賴，以防邊界封閉或認知遮蔽，並能根據下游反饋及時糾偏。但在這一切之前，一個更重要的問題需要得到真正的反思：當人工智能奇點即將降臨的時候，人類是否已經做到真正認識自身了呢？

---

①⑨Upol Ehsan & Mark O. Riedl, “Explainability Pitfalls: Beyond Dark Pattern in Explainable AI”, https://arxiv.org/abs/2109.12480

②⑧Nurhafihz Noor, Sally Rao Hill & Indrit Troshani, Artificial Intelligence Service Agents: Role of Parasocial relationship, *Journal of Computer Information Systems*, 2022, 62: 5: 1009-1023.

③保羅・維爾諾：《諸眾的語法：當代生活方式的分析》，董必成譯，北京：商務印書館，2017 年，第 83 頁。

④吳冠軍：《從元宇宙到量子現實——邁向後人類主義政治本體論》，北京：中信出版集團，2023 年，第 256 頁。

⑤Slavoj Zizek, *Event: Philosophy in Transit*, Lodon: Penguin, 2014, p. 119.

⑥路德維希・維特根斯坦：《邏輯哲學論》，賀紹甲譯，北京：商務印書館，1996 年，第 32 頁。

⑦Ilia Shumailov et al., “The Curse of Recursion: Training on Generated Data Makes Models Forget”, https://arxiv.org/abs/2305.17493

⑩Hans de Bruijn, Martijn Warnier & Marijn Janssen, The perils and pitfalls of explainable AI: Strategies for explaining algorithmic decision-making, *Government Information Quarterly*, 2022: 39:2: 101666.

⑪⑭⑮德勒茲、加塔利：《資本主義與精神分裂（卷 2）：千高原》，姜宇輝譯，上海：上海書店出版社，2010 年，第 295 頁；第 574 頁；第 668 頁。

⑫弗雷德里克・詹姆遜：《新馬克思主義》，陳永國、胡亞敏等譯，北京：中國人民大學出版社，2018 年，第 322 頁。

⑬藍江：《ChatGPT 是否會吞噬我們的剩餘快感——人工智能時代的病理學分析》，武漢：《武漢大學學報》，2023 年第 4 期。

⑯Roberto Finelli, Marx, Spinoza and the New Technologies, *International Gramsci Journal*, 2021, 4:2: 3-24.

作者簡介：吳靜，南京師範大學哲學系教授、博士生導师，南京師範大學數字與人文研究中心主任。南京　210023

[**責任編輯　劉澤生**]