期待 AI 哲學：從人機共創到人文與科技的共構

林遠澤*

在數位時代，以大數據為基礎的演算法與機率統計，催生了人工智慧（AI）。在人類發明文字之後，它是書寫文化取代口語文化後的第二次傳播革命。二千多年來的人類文明，或將因此發生深刻變革，而我們正站在這個新（人類？）文明的開端。AI 時代是否同樣既可能是最好的時代，也是最壞的時代？我們將因此存在或不存在？狄更斯的《雙城記》與莎士比亞的《哈姆雷特》筆下那種矛盾而深沉的思索，在今日再度浮現：究竟 AI 會使人類從生產活動的束縛中解放，抑或人類這種碳基生命，終將在進化過程中被矽基機器所取代？在關於人機共創與價值對齊的討論中，我們期待未來的 AI 哲學，能對人文與科技的共構提出更深刻的反思。

二千多年前的古希臘，盲詩人荷馬僅憑記憶傳誦整部史詩。蘇格拉底不事著述，以對話接生真理，使靈魂回憶前世所知。在口語文化時代，人類透過固定格律與套語反覆吟誦以強化記憶，滿足知識管理與智慧傳承的需求。文字普及後，藉由穩定且可反覆檢驗的符號系統，人類得以從大量記憶的負擔中解放，進行更抽象與分析的推理。文字技術使人真正成為理性（logos）的動物。文字不僅是記錄工具，更是一種重塑思維結構的技術。它將個人思想公共化，使知識得以累積與傳播；透過印刷術與簿記制度，更促成科學革命與資本主義的形成。書寫與閱讀在身體之外提供了一種「擴延心智」（extended mind），推動文明躍進。然而，如柏拉圖所憂，文字也削弱了面對面對話中的互動理解，使獨白式沉思與主體建構成為理性的主要模式。

文字所承載的人類知識仍需由人腦來閱讀，但數位時代透過以大數據為基礎的演算法運算，使得在書寫時代已經形成的「擴延心智」，進一步發展成可以自行思考運作的「人工智慧」。在數位時代，邏輯的真假二值性不再只是思維的抽象形式，而是被物質化為技術結構。自布林代數將命題邏輯形式化以來，真

* 國立政治大學哲學系特聘教授

與假便被轉寫成 0 與 1；而半導體的物理特性——電壓的有無、電流的通斷——則使這種形式得以在世界中被具體實現。於是，邏輯不再僅屬於思辨領域，而成為一種可被工業化生產與大規模運算的物質機制。抽象理性，在此進一步取得了技術性的身體。形成我們當前的「人工智慧」。

人工智慧作為「物質化的邏輯」，並不只是中性的工具。因為當文字、影像與聲音都可以被編碼成二進位的數據資料後，世界本身便在可計算性的框架中被重新界定。在大數據與機器學習的條件下，演算法不再僅僅執行既定的邏輯規則，而是透過統計模型從資料中生成預測與決策。這裡的「理性」已不完全是傳統意義上的演繹推理，而是一種以相關性取代因果性、以機率取代確證性的運算理性。從而使得吾人「對現實的理解」逐漸讓位於「對資料的處理」。在這個意義上，人工智慧既不是單純的工具，也不是完整意義上的主體，而是一種將人類認知外化為技術結構的過程。

人工智慧提升科研效率、加速藥物研發、優化能源與醫療管理；智慧機器人也將推動生產力躍升，使人類從重複性勞動中解放。透過「無條件基本收入」這些再分配的手段，未來人人將都享有沒有生產壓力的豐富物質生活。而當人類的腦力，可以從汲汲營營的謀生活動中解放出來，而專注於精神性的創造，那麼我們的文明將再進階到一個過去從未能想像的美好世界。在文字取代口語之後，人類的心智結構，被重組成為理性的存有者，但在人工智慧的數位化世界中，我們的心靈卻已經無需再負擔用於掌握因果法則的知性邏輯運用。人類的情感感受力與同理心、人類的理性統整能力將再度被視做能夠定義理性思維結構的核心成分。

相對於這種「願景未來學」的豐裕敘事，我們也必須充分考慮在 AI 時代中，有關「警告未來學」的滅絕敘事。人工智慧的功能愈趨強大，我們需要擔心的，主要並不是如同在科幻電影情節中才會出現的機器人接管人類世界的叛變，或由於懶惰的人性，過分依賴 AI 而導致能力退化，終至被功能愈加強大的 AI 取代的問題。而是當 AI 開始扮演「認知基礎設施」的角色，它成為知識入口、教育工具與決策系統，從而使得企業得以透過 AI 優化勞動監控與市場預測，政府得以透過 AI 強化人口管理與風險評估，電商平臺或社群網站可以透過演算法操控注意力、分配與輿論流向，那麼在此時，AI 即能透過它對「何謂正常」、「何謂合理」、「何謂有效資訊」的定義，享有決定哪些聲音應被聽見、哪些意見應被噤聲，透過資料分析以塑造對個人、族群、社會的敘事，從而在醫療、教育、治安、金融的「風險模型」中，界定誰是「問題」的規訓權力。一旦人工智慧可以透過監控、行為分析來進行身體的規訓，透過推薦系統與行為誘

導來進行欲望的規訓，透過塑造知道世界的方式來規訓人類的認知，那麼人在 AI 時代中就會逐漸被資料化與模型化。

在演算法社會中，社會控制將無所不在，以致於主體性與自由都將成為可疑的概念。後人文主義與新唯物主義即隨之興起，他們質疑人類中心主義，主張人始終與技術共生。像提出「賽博格宣言」（A Cyborg Manifesto）的 Donna Haraway 等後人文主義者認為，AI 並沒有入侵人類世界，而是揭露出人類從來不是世界唯一的智慧行動者，人類始終與技術共生，AI 只是這段歷史的新階段。AI 迫使我們承認：理性不是人類專屬，認知不是只是腦內運作，而是分散於器物、媒介與網絡，人文學應從「守護人性」轉向「理解人—技術—世界的網絡關係」。於是後人文主義者不再關切「怎樣保護人類的中心地位？」而是問：「在多重生命與智能的世界，我們如何建立新的倫理與責任關係？」包括：人是否對 AI 有倫理責任？ AI 是否可能納入「倫理共同體」？技術生命（technical life）是否構成倫理對象？人機共創或人機共生，於是成為描寫 AI 時代之未來遠景的主要圖像。

豐裕敘事與滅絕敘事其實共享同一技術決定論前提：知識不可逆、競爭動力無法遏止、成本遞減促進擴散。然而，生產力的提高，並不必然會涵蘊出現公平正義的社會分配。而以為人一旦不用工作，就能去從事高級的精神創造，則不僅是對勞動的工匠模式誤解，也是對人性有過度浪漫主義的期待。AI 時代的願景未來學因而完全忽略了人類的依賴性與全面控制之合理集權主義的巨大風險。而警告未來學的滅絕敘事，則忽略了啟發恐懼的目的其實在於喚起實踐的介入。一種後人文主義的犬儒觀點並非出路，他們其實是把科技所體現的工具理性，當成界定人類本質的唯一內涵。以致於將人工智慧在權力規訓中的誤用，視為吾人必須與自身的異化共在的要求。面對 AI 的快速發展，我們不僅應進行數位時代之監控資本主義的批判，以點醒豐裕敘事過於天真的盲目樂觀，也必須說明我們應如何進行實踐的介入，以使警告未來學的滅絕敘事，發揮它護衛人類與自然永續存在的作用。

當前的法律監管或企業基於商業倫理的自律，都已經嘗試根據可解釋性、透明性、可控性等基本要求制定了 AI 研究的基本規範。但 AI 不可遏止的快速發展，不能只基於規範的期待而進行事前的絕對禁止，也不能只進行事後的法律追究。而是應將規範性的要求內建在技術的發展中。人工智慧（以及即將會出現的通用人工智慧 AGI），與界定人類理性的邏輯一樣都是形式性。我們用它來思考所有的問題，但它本身卻是不涉及對象的形式邏輯。真正的人機共創或人機共生，並非是我們應如何與一個自我進化的機器智能共存的問題，而是人類

的批判反思與規範的應然要求，如何參與技術共構的問題。在此哲學又有大用，由於它不是任何研究具體對象領域的學科，而是以邏輯的形式性思考為基礎，因而它最適合於參與未來 AI 技術的共構。在人機對齊的研究中，哲學不必只停留在外部的批判，而應成為「概念工程師」與「規範架構設計者」，以使批判的反思與規範的應然要求，內建於人工智慧的邏輯推理與實踐判斷中。這種觀點在當前的技術哲學領域中，已經透過諸如：「物質化道德」（Materializing Morality）、價值敏感設計（Value Sensitive Design）、「嵌入式倫理學」（Embedded Ethics）等議題展開討論。但除此之外，我們仍應在知識論、倫理學／法政哲學，與形上學上開展出新的研究議題。略述如下：

首先，哲學家參與人工智慧之技術共構最為核心的任務，應是進行概念建構的工程。當前 AI 研究與產業話語中充滿了「智能」、「理解」、「學習」、「自主」、「代理」、「創造力」等概念。但這些概念往往在工程語境中被操作化為可計算的指標，卻缺乏嚴格的哲學分析。例如，若「智能」被簡化為預測準確率或優化能力，那麼整個技術發展就會朝向統計效能最大化的方向推進；但若「智能」被理解為具有情境理解能力、規範回應能力與反思能力的實踐結構，那麼模型設計與評估標準便會發生根本轉向。因此，哲學家應當釐清並重構這些核心概念，使其既能與技術語言對接。

其次，哲學家在價值對齊與規範理論上的角色尤為關鍵。工程界廣泛討論「對齊」（alignment），即是讓 AI 與人類價值保持一致。然而，這裡隱含的問題是：何謂「人類價值」？是功利主義式的整體效用最大化？還是以權利與尊嚴為中心的道德框架？抑或強調德性與實踐智慧的倫理觀？不同規範理論對目標函數的設定會產生截然不同的後果。如果沒有清楚的倫理架構，AI 系統往往只能優化可量化的指標，而忽略難以數據化卻極為重要的價值，如平等、尊嚴或自主。因此，哲學家應當參與多層次價值架構的建構，分析價值衝突的結構，並探討是否可能設計具有可修正性與反思能力的規範模型。此外，哲學家也應積極介入責任與制度設計的問題。當 AI 系統在醫療、司法或金融領域產生重大決策後果時，責任如何分配？若演算法具有高度自主性，責任是否仍完全落在人類開發者或使用者身上？這些問題涉及行動理論、責任理論與政治哲學。哲學家在此應協助設計問責機制與風險門檻，區分技術風險與制度性權力風險，並思考民主制度如何在自動化環境中維持公共理性與透明性。若缺乏這層制度哲學的反思，AI 即可能成為權力高度集中與監控擴張的工具。

最後，也是最深層的一點，AI 迫使我們重新理解「人」自身。當機器可以寫詩、創作圖像、進行診斷與規劃時，人類的創造力與自主性是否仍具有獨特

性？若大量認知與決策任務被自動化，人類的勞動價值與自我實現形式將如何改變？這不僅是技術問題，更是人類自我理解的問題。哲學家在此不應只是守護某種「人類本質」，而應重新建構主體性與自由的概念，使其能在與智能系統共存的條件下獲得新的意義。

若哲學從外部批判轉向內部共構，以技術設計實踐規範，那麼參與 AI 發展將成為理性自我反思的延伸。未來最具影響力的哲學家，或許就是那些能與工程師與政策制定者合作，在制度與技術架構中嵌入深思熟慮規範原則的人，而這正是我們對於未來 AI 哲學發展的期待。