機器-人-明日社會

「人工智慧、心靈與演算法社會」討論會紀實

作者：蘇上雅[1]

前言

華人社會有句俗諺：知人、知面，不知心。人類心靈之多變難料，彷彿從古以來，跨文化傳統社會皆有目共睹，也一直是各領域學科研究企圖攻破的難題。然而，上述基調，卻可能不是定調，尤其在「人工智慧（Artificial Intelligence，下簡稱 AI）」問世之後。從屏幕上一部部預示「機器人」與明日社會的科幻片，到當代實踐上屢屢突破且逐步深入人類生活層面的人工智慧技術，人類心靈的複雜謎題未解，另一個課題（憂慮）卻已迎頭趕上：人類的心靈是否有被「人工智慧」仿效、參透、甚至超越的可能？當人工智慧進入人類社會，人類社會互動之方式與規範將如何改變？

6 月 8 日由臺灣法實證資料庫、臺灣法學會、臺灣法理學會、臺大基礎法學研究中心合辦之「人工智慧、心靈與演算法社會」論文討論會，由建置臺灣法實證研究資料庫第五期計畫主持人、臺大法律系陳昭如教授主持，由臺大法律系顏厥安教授主講，並邀請臺大資訊網路與多媒體研究所林守德教授、中正大學哲學研究所謝世民教授，以及中研院法律學研究所博士後研究員陳弘儒博士一同與談，期藉由資訊、法律、哲學等多元領域的對話，一起思考當代便利的資訊生活中，逐步突顯的「AI 倫理學」議題。

並不是 AI 越來越像人，而是人類社會越來越 AI！

討論會一開始，主講人顏厥安教授先以電影《2001 年太空漫遊》中「墨闇碑（monolith）」之象徵起手，談及心靈之客觀化，指出：人類心靈的運作，雖然有晦暗不明、無法被觀察的面向，但亦存在外顯、客觀、展演的面向。轉而談到 AI，主講人以「會下棋的 AI」AlphaGo 系列的突破為例，說明下棋程式的演算法基本特色，指出 Al 學習的速度很快，不需要人類指導，只需要告訴它規則，AI 就會自己學。

主講人進一步由 AI 倫理學的發展出發，談及對 AI 前景的三種態度：樂觀論、悲觀論與懷疑論。樂觀論認為，不論是機器能力的發展，以及對人類社會的貢獻，都可以抱持著正面肯

[1] 臺灣大學科際整合法律研究所碩士。

定的態度；悲觀論認為，機器/AI 的潛力很大，能力會越來越強，但是對人類文明的威脅也不斷增加，甚至未來可能出現由機器/AI 主宰的狀況；懷疑論則認為，AI 受限於其基本的運算能力特色，永遠不可能發展出類似或接近於人類的心靈或智慧能力。主講人從悲觀論出發，但回頭反思 AI 倫理學的「機器人倫理」三法則「不傷害、服從、保護自身」，認為該理論聚焦於「人性」，忽略了社會結構、關係面向。主講人認為，AI 倫理學的首要重點，並不在於擔心 AI 是否能夠產生高度接近，甚至完全相同於人類「心靈」的能力；而是人類社會的運作，是否越來越依賴演算法進行，以至於社會運作漸走向「演算法社會」(Algorithm Society)？易言之，重點並不在擔心 AI 越來越像人，而是人類社會越來越 AI！

最後，主講人以法理學「內在觀點」概念切入，討論 AI/演算法有無內在觀點；並進一步指出：AI 雖無內在觀點、沒有能力透過自我認知形成一套社會規則；但在演算法社會中，不具備對規範之內在觀點與自主意識的演算法，正逐漸取得對人類社會的掌控權。

工程觀點，談 AI 的侷限與潛能

第一位與談人林守德教授，從資訊工程的角度出發，談及資工領域研發 AI 的歷史、AI 的侷限與能力。林教授提及，工程上對 AI 的定位分成兩類：strong AI、weak AI。strong AI 論述始於 70 年代，企圖用相同於人類產生智慧的方式培育 AI，後來失敗，原因是：人類產生智慧的機制實在太複雜。90 年代另一批研究者提出悖論：人做不好的事情，AI 做得好；有些 AI 做不好的東西，人卻很擅長。舉「下棋」為例，AI 很擅長，但對人來說，下棋卻是亟需智慧的活動；反之，若是「人臉辨識」，人三歲就會了，AI 卻直到這兩年才趕上。

AI 究竟擅長什麼？「選擇題，」林教授指出，AI 的整個機制，使它只能 output 一個選項，關於 AI 的技術不過如此。但，為什麼 AI「看起來」那麼有智慧？那是因為：很多事物，都能夠變成選擇題。包括預測明天股市會不會漲、預測明天 PM2.5 的值，甚至是寫詩：對人類來說需要高心靈智慧的創作活動，對 AI 來說，就是一道接著一道的選擇題從它看過的萬句詩中不斷延伸選字的過程。AI 的「選擇」可以呼應主講人「區隔」的概念，它透過這樣的能力去組合出讓人看起來很有智慧的東西。

不過，林教授也提到 AI 的侷限，乃是「自我意識」。現在的技術，無法讓電腦擁有自我意識，其理由亦在於：自我意識不是選擇題。自我意識，意味著它要瞭解到自己是個 AI、自己是一個東西，要能夠根據「它的意願」去做事情。對電腦來說，它只能去「選」，它無法透過選擇自己是人還是 AI 這件事情，「自我意識到自己是 AI」。進一步延伸：沒有自我意識的東西，是否算是心靈？AI 外顯看起來可以做到很智慧的東西，但 AI 解的東西，都是人預設指定、給出選項（predefine）的。AI 並沒有在此之外的能力。

換言之，即便是看起來可以自己下棋的 AlphaGo、AlphaZero，它「學習」的過程，要嘛是看人類過去的選擇來選，或是告訴它這樣下得到的結果分數加分或扣分，進而模擬學會的。相對於下棋等 game 性質的事物，可以透過模擬過程來達到；日常生活事物，因為現實生活中的經驗很難透過模擬來達成。因此，AI 處理的東西，目前大多還是侷限在可以一再事先模擬的事物。

最後談到 AI 與社會的關係，林守德教授指出，目前此議題在工程領域尚非主流，目前資工研究的主力仍放在：如何使 AI 更聰明、讓它不只會做選擇題。舉例來說，「證明題」、「申論」需求產生邏輯論述，很難被模式化為選擇題，AI 目前做不到。反之像詩，沒有邏輯、可以跳躍，沒有邏輯你會覺得是自己理解上的問題、沒有邏輯反而有意境，人不會覺得是 AI 出錯。

法學觀點，談演算法社會的擔憂與可能

第二位與談人陳弘儒博士，從基礎法學的角度出發，回應主講人兩個大哉問：什麼是演算法社會的深層擔憂？以及，AI 是否有「內在觀點」可能？

於第一個議題，即「演算法社會最深層的擔憂是什麼？」陳博士肯認主講人的基本界定，認為人類社會確實將越來越朝向 AI 演算法的運作方式邁進，而心靈的創造性、美感等能力則可能被邊緣化、不受重視。但陳博士進一步指出，演算法的應用在人類社會由來已久。譬如食譜，讓人透過簡單、有限的步驟，達成完成一道菜的任務，就是一古老的演算法應用。陳博士認為，因此，值得更進一步探討之議題，應在於：演算法的應用歷史如此久遠，為什麼它「現在」變得那麼重要？

再者，陳博士認為，「究竟何謂人工智慧？」亦是需要先行界定的概念。陳博士偏向從「代理人」的概念來界定 AI，亦即：原本人可以自己做的事情，「外包」給另外一個對象來做。陳博士認為，在 AI 代理人類社會生活的課題中，除了傳統法學界已論及的「代理權」、「法律關係」議題外，另一值得深思的問題是：AI 的能力，是否足以作為「代理人」，亦即 AI 是否具有理解人類的語言、能夠推論等等能力？

由上述議題，陳博士進一步延伸出一重要論題：是否人類心靈的所有能力，都能夠用演算法呈現出來？（諸如人類的情感表示、理性推論或決定等）並進而談到心靈哲學領域所提出、值得深思的主張：「軟體：硬體＝心靈：大腦」。從上述等式進一步推伸，陳博士論及 AI 研究對探索心靈的可能助益：若上述等式成立，縱使不知道心靈實際上如何運作也無妨，透過設計出人工智慧中適合呈現心靈運作的演算法，透過電腦設計演算法，相當程度地，可以去觀看心靈是如何運作、判斷的。陳弘儒博士認為，換言之，演算法之發展或有一堅實理由在於：完

備人類心靈、完備呈現心靈原有的能力。不過，陳博士於此打住，並談到需要同步思考的是，是否有朝一日，AI 的運作雖是透明的，演算法背後的運作邏輯卻無法再被人類觀察、理解？

關於「AI 是否可能有規範性/內在觀點？」一題，陳博士肯認主講人的主張：AI 沒有內在觀點、沒有辦法去理解/接受所謂社會規則之意思，因此無法產生規範性。但陳博士認為，值得進一步思考者是：在觀看 AI 的運作時，所謂「規範性」的意涵，是否仍要限縮在傳統的規範性概念？以人臉辨識為例，陳博士指出，雖然 AI 無法以人的方式或標準，去瞭解/認知不同外表特徵的「意義」為何，但它確實可能透過再現特徵、重複套用到具體案例上，去做人臉特徵差異的「辨識」、進行概念的運用。陳博士認為，從 AI 的學習、辨識，值得思考所謂「沒有意義的概念」：在當代概念的運作中，把意義的理解取消，可能已非重點。而若是如此，陳博士認為在當代社會，規範性的客體也可能產生轉變：過去也許必須將實質理由、意義呈現出來，方能作為自己行為或批判他人行為的標準；當代社會生活中，行為背後的意義需求強度或減弱，不需要理解理由的規範性。陳博士最後以「小孩習得排隊」生動舉例，說明所謂弱意義的規範習得過程：小孩被教導應該排隊時，其不見得「理解」其中意義，但小孩會去「觀察」其中的「規律的運作」：應該要排隊，因為大家都排隊。進而小孩習得排隊的規範。

哲學觀點，談演算法社會的未來

第三位與談人謝世民教授則從哲學觀點切入，回應今日主講、討論的議題。謝教授首先談及：在哲學領域，以往 AI 倫理學較受重視的議題，較從「個人」角度出發，論及 AI 的選擇，可能牽涉影響的個人安全問題，與主講人關注的「演算法社會」，抑或較宏觀的人與人之間的「關係」、「社會性元素」，有所區別。

再者，謝教授亦回應 AI 的「規範性」議題。謝教授總結前述主講人與兩位與談人對 AI 無自我認知能力的討論，進一步論及：AI 雖然不具有「主觀規範性」、沒有主觀上接受規範拘束之可能，但 AI 可能在沒有主觀認知的情況下，去「遵守」演算法的指令。謝教授援引拉茲提出的另一種規範性概念：「分離觀點（detached point of view）」，亦即在對規範有距離、沒有意願、甚至不涉及意願的前提下，仍可能遵守規範的運行。謝教授認為 AI 的運作即是「分離觀點」的具現：某程度上演算法決定 AI 的「選擇」，但 AI 其實是在沒有內在觀點認知的前提下在執行演算法的規範。