從人類與 AI 的主奴辯證到 X.A.I.

林從一*

人類與AI的主奴辯證

科幻小說、動漫和電影（如《銀翼殺手 2049》）常出現人類／AI 主奴辯證，喚醒人類的深層恐懼：人類創造 AI（人工智慧）[1]，AI 服務人類，服務人類要瞭解人類，做好服務要能解決問題、評估風險、進行預測、採取行動，而在可見的未來，平常如臉部辨識，複雜如倫理決定或新病毒基因體定序，AI 似乎都會做得超乎人類能想像的好、超乎人類能理解的好，但是，享受 AI 服務要付出終極代價。

科技的發展讓人們越來越相信，AI 可以比你還瞭解你的喜好、AI 可以比你還瞭解你的喜怒哀樂各種情緒、AI 可以比你還知道你的價值傾向、AI 可以遠遠比你知道這世界發生了什麼事、萬事萬物之間如何關聯，因此，AI 可以比你知道如何設定目標、解決問題，從而提出更合你意的建議，並且能採取比你更即時、更有效的行動。因此，人類似乎不僅「終究會」而且也「理應該」把選擇權、決定權、行動權等等這些標誌人性的自主權讓渡給 AI，最終，在演化的進程中，AI 甚至也應該取代人類這種生命型態。人類與 AI 的主奴辯證終局不是 AI 這個奴隸翻轉成為人類的主人，因為，AI 將不需要人類，人類可以做到的 AI 都可以做得更好。

人們已經開始利用 AI 進行量刑建議、疾病診斷、借貸風險判斷等等影響人的生活與生命的決策，既然影響人的重大權益，該負責把關並負最後責任的是人類，而不是 AI 系統。AI 系統如何影響人類決策，人類應該如何管控 AI，都是應該研究的主題，但是人類／AI 的主奴辯證困局的解方不在「控管」，而在於讓 AI 成為我們人類的一分子。

* 國立成功大學通識教育中心教授、人文社會科學中心主任、副校長

[1] 本文的 AI 專指透過張量計算來處理非結構化數據的次符號模型，簡單地說就是連結主義式（connectionism）的 AI，而非處理命題（如信念和知識）的符號模型的 AI。

X.A.I.——打開AI黑盒子

人類擅長的或自我意識到的認知模式是一種處理命題、信念、知識的線性、遵循有限規則的符號運算過程。AI 則是處理數據的次符號（sub-symbolic）模型，它的表徵結構是具多隱藏層（深度）的離散系統，也就是連結主義式（connectionism）的認知模式。AI 的運算功能可以處理海量數據，遠比人類總加可以處理的還多得多，而更關鍵的是，針對我們餵食給 AI 系統的海量數據，AI 系統會創造出極為複雜且高度非線性的內部表徵結構。有多複雜？AI 模型可以有 3 階或更多階的互動網絡，且現在市場上表現不錯的 AI 模型的內部表徵結構都涉及約 1 億個參數，換句話說，那些 AI 系統所做出任何一個決定，譬如判斷某一張圖片是「這是一隻拉不拉多犬」或「這是一個玩抓娃娃機的老男人」，都涉及 1 億個數字的組合。正是那麼巨大的參數，讓 AI 可以注意、計算無窮小的訊息，衡量人類無法衡量的訊息，從而發展出人類無從探知的、看待、思考事物的全新方式，而這正是 AI 令人驚豔的地方。

簡單地說，海量的數據、非線性的表徵模型讓 AI 可以做得比人類更好、更快、更強，而越厲害的 AI 系統越是個黑盒子（black box），我們越無法瞭解它們是如何做出決定的，無從理解 AI 的決策所根據的理由是什麼，因此也無從探究、無從質疑、無從評估、無從問責。

人類是理性動物，人類社會是理性社會，人類活動的特徵就是「給理由也要求理由」（giving and asking for reasons）。理性、說明（可理解、可詮釋）、意義、責任、信任、行動、自主這些概念彼此緊密關聯，它們共同構成人與人類社會的基本特徵。因此，當人類越享受黑盒子 AI 的服務時，越依靠 AI 的建議、選擇而行動、過日子時，人類就越無法理解他們自己為何做出這些行動、為何那樣的生活。讓渡出選擇權之後，就讓渡出可理解性，人的面貌就逐漸消失了。

在生活的方方面面上，我們讓 AI 幫我們做出越來越多重要的決定，對人的福祉影響重大的決定，我們需要瞭解 AI 是如何做出那些決定的，我們需要「AI 的決策能被說明」，AI 的決定策略必須透明化。這個議題及研究領域稱為 X.A.I.，X.A.I. 是 Explainable AI（可理解的人工智慧）的簡稱。X.A.I. 是一個艱鉅的挑戰，AI 之所以神奇，正是因為 AI 以非人的方式認知、思考世界，從而產生全新的認知、思考模式，進而揭露人類無法揭露的事實，「以人類的方式認知、思考」（讓人類能夠理解）與「全新且強大的認知、思考」似乎不能兩全。

所幸，X.A.I. 還是有些進展，而 X.A.I. 的議題其實就指出了它的研發策略，AI 的可理解性必須是一個同時結合人類認知模式與 AI 認知模式的系統，也就

是一個同時立基於（次符號的）數據與（符號）命題（知識）的認知系統，亦即，X.A.I. 的主要課題是將兩種表徵系統結合起來：一種是常用在處理文本和語言等符號性對象的語意符號空間（Semantic Symbolic Space），另一種是常用在處理圖像和語音等次符號數據的特性向量空間（Feature Vector Space）；從而形成一個新形態的空間，稱為語意向量空間（Semantic Vector Space）。語意向量空間建構可以有兩個方向：一是將語意符號空間嵌入（embedding）特性向量空間，也就將符號和符號之間的關係轉換成特定的向量，舉例來說，這樣的方法可以讓我們以視覺化的方式瞭解「拉不拉多犬」與「貴賓犬」這兩個詞彼此之間的關係（其實，這也是一種讓 AI 聽得懂人話的方法）；另一個方向是將特性向量空間進行語意上揚（raising）的處理，使得處理次符號數據（如圖像）的隱藏層也可以獲得語意性質，這可以讓我們得以系統性的整合圖像、行動、文字和語言。而重要的是，特性向量空間的語意上揚使得 AI 的決策黑盒子被打開，使得 AI 可以說明自己如何做決策，從而使得 AI 能夠被問責。

X.A.I. 有不同的研究策略，其中一種涉及深度神經網絡（deep neural network, DNN）的模式。[2] 在傳統的圖像、行動識別的深度神經網絡模型中，輸入海量圖片、動作之後，我們會得到文字描述的結果，但是我們不知道隱藏層（hidden layers）所歸結出的中介特徵是什麼，我們也就不知道為什麼會得到這樣的結果，而如果最終的文字描述判斷是錯的，我們也不知道為什麼錯。如何進行語意上揚工作，打開 AI 的決策黑盒子？

有一種策略是讓 DNN 在建構次符號數據的特性向量空間時，也同時讓 DNN 學習「人類的認知模式」，也就是餵養 DNN 相關語意訊息數據，形成語意向量空間。舉例來說，我們先獲得人類對圖片描述的語意訊息數據，再將這些語意數據和圖片同時餵養給 DNN，這樣我們就可以知道 DNN 中隱藏層所涉及的中介特徵是什麼，因為每個「神經元」都會因此帶有語意訊息數據，整個網絡因此也變成可解釋。簡單的說，透過讓 AI 學習表徵一組對象的同時，也學習「人類是如何以語言描述同一組對象的」，使得 AI 不僅可以描述圖片中的各種對象，也可以說明它是如何做出這些描述的，換句話說，AI 因此可以用人話來描述自己的決策與過程和理由。

[2] Yinpeng Dong, Hang Su, Jun Zhu, Bo Zhang (2017). Improving Interpretability of Deep Neural Networks with Semantic Information, 10.1109/CVPR.2017.110.
Jindong Gu, Volker Tresp (2019). Semantics for Global and Local Interpretation of Deep Neural Networks, arXiv:1910.09085, 10/21/2019

AI成為我們的一分子

當 AI 可以說明它們自己是如何做出決定的，人類因此可以瞭解 AI 決策所根據的理由是什麼，那麼，無論是 AI 的決策或是人類藉由 AI 所做的決策，都因此可以加以探究、質疑、評估與問責。而當 AI 不僅可以說人話解釋它自己，同時也可以聽得懂人話，那麼人類與 AI 之間的真正溝通就開始了，它開始瞭解人類社會中細緻的分寸，它瞭解它不能只想贏或成功，它知道別人的成功有時候就是它自己的成功，它知道自己的限制是什麼，它知道它也需要人類的協助，簡言之，這時候，AI 在人類社會中會獲得某種行動者、主體地位（agency）。

如果 AI 是不可避免的，能夠與人類溝通、一起合作的 AI，遠比黑盒子 AI 讓人類信任，而人類與 X.A.I 的合作才能創造未來的倫理環境。

一個內建人類觀點的 AI，才是一個可理解的 AI，一個能夠以人類語言說明自己並與人類溝通的 AI，才是一個可理解的 AI，而屆時，AI 也將獲得人類社會中的主體身分，成為「我們人類」的一分子。更核心的是，當 AI 與人類成為「我們」，彼此之間的倫理關係將徹底改變，不再是主奴辯證中主奴的倫理關係，人類被 AI 取代的危機感才會消解。