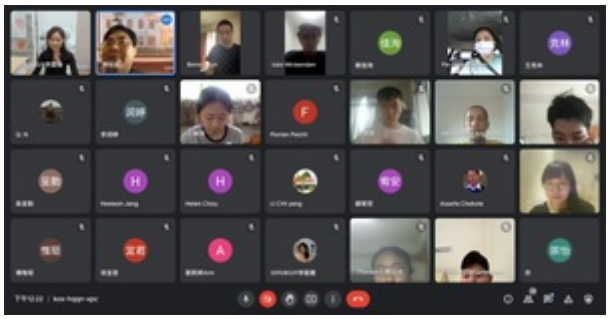
創國學院AI與治理演講　李韶曼談人工智慧決策成形機制

字體大小調整 小 大

日期 : 2022-06-09　單位 : 創新國際學院



由創新國際學院暨法學院助理教授陳柏良主持，邀請國立成功大學敏求智慧運算學院助理教授李韶曼演講，和學生分享人工智慧決策成形的機制（照片來源：創新國際學院）

【創新國際學院】

本校創新國際學院（創國學院）、人文社會與科技前瞻人才培育計畫「『心．機』共融」計畫、本校台灣研究中心於5月24日共同舉辦「可解釋的人工智慧」（Explainable AI）英語講座，由本院暨法學院助理教授陳柏良主持，邀請國立成功大學敏求智慧運算學院助理教授李韶曼演講，和學生分享人工智慧決策成形的機制。

本講座以線上的方式進行。首先李韶曼從State v. Loomis的案件談到AI技術應用的爭議，以及AI自動化決策中所產生的偏見與歧視。李韶曼提到，評估罪犯再犯風險的人工智慧系統COMPAS，在美國法庭上常用來建議刑度。既有研究也提到，COMPAS系統預測再犯率的準確性僅只有60-70%左右，其中黑人再犯分數顯著高於白人再犯分數，若對照實際再犯率，黑人更可能受到不利的預測。實際上運用於訓練AI模型的資料是社會產物，若我們無法看見其中隱含的差別待遇，自動化決策自然難以如我們想像的衡平客觀。

講者李韶曼認為可解釋性AI的意義在於作為使用AI，部署AI，研發AI，甚至管制AI的基礎（照片來源：創新國際學院）

李韶曼接著以深度學習的醫療影像應用為例，學者發現AI竟然可以從X-Ray影像中辨別患者的種族，甚至達到90%的準確率，目前原因仍有待確認。李韶曼指出這可能與X-Ray影像的後設資料與潛在資料相關。複雜的人工智慧模型可以從各種數據資料，產生人類難以預料、無法解釋的成果。同樣地，我們也很難辨別人工智慧模型產出的結果是否有誤，亦很難解釋結果。我們該如何確保人工智慧系統能夠符合研發或部署的目標？

在難以完全確保AI自動決策的結果正確與否時，可能造成使用者的信任危機。比如2021年，美國威斯康辛州Kenosha County Circuit Court在審理槍擊案件時，即有是否能讓陪審團使用iPad zoom in影像功能來檢視證據的爭議。被告律師認為，iPad所使用的人工智慧技術，不僅只是強化影像，更可能是在創造場景細節。在檢方未能進一步舉證說明前，法官令陪審團僅能以原監視器輸出影像來審酌。

這些事件帶領我們反省，為確保AI的使用結果合乎預期，讓AI公平、可靠、進而讓使用者信任，對AI透明性與可解釋性的探究，知悉AI系統所使用的資料、演算法運作的因果邏輯，甚至是反事實解釋，或許是必要的第一步。當前可解釋性AI研究，不但與以往社會 - 技術（sociotechnics）研究中的幾個主要議題Fair Machine Learning, Human-in-the-Loop Autonomy和AI Safety外，更是跨學科的重要議題，不同領域都應該參與討論。

可解釋AI理念廣見於當前AI倫理規範中，但要付諸於現實，需要規範制度、軟體、硬體三個層面的配合。規範面而言，針對公部門所部署的AI決策系統，民眾有知的權利，這需要政府進一步提供配套措施，方能完整保障。而私部門所部署的AI決策系統，使用之資料、演算法僅為少數人掌控，一般使用者難以知悉。歐盟近年提出GDPR（一般資料保護規則），為人工智慧系統應用提供個人隱私和資料保護之框架，其中是否包含使用者請求解釋AI自動化決策的權利，仍有爭議。依照歐盟Data Protection Directive下成立之諮詢性組織the Article 29 Working Party的觀點，GDPR 第13條至第15條，以及第22條，應可作為請求解釋AI決策的權利基礎。

針對軟體部分，已有explainable AI技術之研發，試圖理解AI內部運作機制。不過現行技術無論是針對特定AI模型的解釋，或是一般性的方法，多半是用更多AI來解決AI問題。在形式上，多半透過AI模型的後處理來提供文字及可視化界面，是否真能滿足不同利害關係人的解釋需求，仍有待從具體使用情境中加以檢視。社會心理學及其他各領域專家的參與對話，是未來AI發展的重要關鍵。

李韶曼認為可解釋性AI的意義在於作為使用AI，部署AI，研發AI，甚至管制AI的基礎。最後李韶曼點出應由各利益關係人的角度，在具體情境下考察使用AI與解釋AI的利益與代價，並且可以先由集體權的角度來理解AI的解釋。在科技快速發展下，進一步理解並思考人機權責，方有助維護人機協作社會的基本權利與法治。