AI 與性別論壇

時　　間：108 年 10 月 4 日（五）9:20-12:30
地　　點：國立臺灣大學社會科學院 3F 梁國樹國際會議廳
記　　錄：張澄清（科技部人文社會科學研究中心博士後研究員）

AI與性別：女性主義STS的關切與實踐

引言人：吳嘉苓（國立臺灣大學社會學系教授）

本次論壇邀請國立臺灣大學社會學系吳嘉苓教授以「AI 與性別：女性主義 STS 的關切與實踐」為題進行引言。吳嘉苓指出，AI 常被賦予許多願景，例如自駕車降低意外事故、人臉辨識協助刑案偵察，以及 AI 影像診斷解決醫師過勞等等。然而，人文社會研究也關切 AI 演算法是否可能深化歧視，造成社會不平等。當前對 AI 發展的論辯，與科技與社會研究（Science, Technology and Society, STS）看待科技與社會的觀點相呼應。STS 觀點，建議拋開科技崇拜與科技恐懼的兩極化爭論，而把 AI 看成是不斷演變的一組社會技術網絡，其優劣成敗涉及各種權力的角逐，並有各種的形塑可能性。無論是研發者還是一般大眾，在創新研發或產品使用階段，都可能影響 AI 的發展。吳嘉苓的引言由「AI 的知識生產」、「AI 的使用與政策藍圖」，以及「AI 的再現」等三個層面，探討 AI 的性別政治。內容摘述如下：

一、AI 的知識生產

因資料偏誤而導致演算法失誤的案例頗多。史丹佛大學的資工教授 James Zou 與女性主義歷史學家 Londa Schiebigner 於 2018 年就在 *Nature* 發表評論，提出 AI 可能會造成性別歧視與種族歧視。歧視的源頭在於使用資料缺乏代表性。例如常用於圖像分類之深度學習網絡訓練的 ImageNet 資料庫，共有 1,400 萬張已標註的圖像資料，但超過 45％以上的圖像來自美國（僅占全球總人口數的

本文由張澄清博士記錄整理，經各場次主講人審訂。

4%），合計占全球總人口數 36% 的中國和印度，卻僅貢獻了約 3% 的資料。缺乏多元性的圖像資料，或可解釋為何演算法將著白衣的美國傳統新娘照片辨識為「新娘」或「婚禮」，但北印度新娘的照片則辨別為「表演藝術」或「習俗」。而英語語料庫中陽性代名詞和陰性代名詞的比例不均，差異最大時曾高達 4 比 1，後來自 1960 年代後期的各種平權運動，語言使用有所改革，稱謂的男女代稱才降至 2 比 1 左右。Google 翻譯若奠基於性別不均之語料庫，常在翻譯結果中錯誤使用男性代名詞，有如歧視歷史的復活。至於自然語言處理常使用的詞嵌入（word embedding）或詞向量（word vector）技術，是將英文字詞以高維度的向量來表徵，向量之間的距離被視為語義相似性（semantic similarities）。於是捕捉到「男人」相較於「國王」和「女人」相較於「皇后」，或「男人」相較於「醫生」和「女人」相較於「護理師」等關聯性，則不難想像會出現「男人」相較於「電腦工程師」和「女人」相較於「家管」等具刻板印象的推論結果。

Amazon 於 2014 年為了提高履歷篩選效率，以及避免人為主觀因素導致偏誤，曾測試以 AI 進行履歷篩選的可行性。由於 Amazon 利用公司內部過去 10 年以男性居多的申請資料進行神經網絡訓練，測試結果發現只要出現與女性的字詞則評分偏低，甚至某兩個女子學校的畢業生通過初選的比例都偏低。也就是說，如果有特定群體的較常出現在訓練資料中，則學習演算法會針對此群體進行優化以提升整體準確度，扭曲的樣本資料會因學習迴圈而放大偏誤。原本在教育、招聘、職場等制度環境中，若普遍對特定群體有差別待遇，將促使群體中的人才不斷流失，這類的 AI 篩選履歷若奠基於過往的資料，很可能維持、甚至是強化了原本的管漏現象。

圖一：吳嘉苓教授以「AI 與性別：女性主義 STS 的關切與實踐」為題引言

具有平權意識的科學家也積極想要修正 AI 可能會犯的錯誤。兩位少數族裔的年輕女性資料科學家 Joy Buolawini 與 Timnit Gebru 為修正人臉辨識技術的性別與種族偏誤，就提出新的基準資料庫，深受注目。她們選定 1,270 名非洲與北歐的國會議員照片，確保膚色與性別的光譜能夠具有代表性，就讓日後檢測人臉辨識精確度時，提供更好的基準。她們以新建立的基準，驗證市面上三大人臉辨識技術，也發現各個人臉辨識技術對於淺色男性的判斷最精確，而深色女性的誤差較大。其實，建立基準而進行改正行動，也發生在早期的科技發展歷程，例如汽車碰撞測試多以男性身體為標準，女性則是縮小版的男性，日後才逐漸增加小孩及孕婦作為檢測的假人模型。

整體而言，在性別化創新時的過程中，有許多方法可以去除偏見，例如在選擇學習資料庫時能試著理解建置過程並揭露資料特性及可能缺失、在 AI 設計的過程中引入公平性審查或建立內部除偏見機制、廣納各界共同面對 AI 造成不平等，以及包容多元的知識理解方式，避免以 AI 作為優位獨大的知識系統。

二、AI 的使用與政策藍圖

演算法已經在各種層面使用，也有許多研究者警覺其造成的傷害。數學家 Cathy O’Neil 甚至稱之為數學毀滅性武器，引發諸多討論。O’Neil 提出，數學模型本就有各種運用，像是棒球賽的分析、甚或是家庭煮食的調整，目標明確、過程透明。但是包括運用做某些犯罪預測，模型不透明、形成有害的迴圈，且可擴大應用的規模，則可能成為數學毀滅性武器。以犯罪熱點預測為例，由於社會邊緣化群體的逮捕率與定罪率較高，以此統計資料為基礎進行預測，並進一步加強相關區域的巡邏勤務，將使得先前邊緣化群體更常因些微過錯而遭逮捕，再加上人臉辨識技術對深膚色與女性的偏誤，執行此政策往往強化原有的偏見，有時也過度仰賴 AI 的判斷，排擠了警察巡邏社區所進行的溝通與探查任務。目前有些 AI 運用政策可能同時包含程序不透明及工程師技術的黑箱，佯裝客觀公正，但實際卻有排序或操作，若運用在求職、醫療或犯罪領域，則結果將不被信賴。著眼於 AI 諸多倫理議題，2019 年 3 月就有 78 位科學家聯名，希望 Amazon 能停止販售人臉辨識技術給執法單位。

AI 機器人介入照護工作是另一個與日常生活有關的重要應用領域，但能否解決現有的問題則仍待商榷。由於照護涉及了勞動付出、情感投入及倫理政治等面向，因此照護型機器人介入社會的評估，有多重層面：機器人能否替代或補充所需的照護勞力？促成的感情交流是值得認可的新型態、還是間接提高情

緒勞動的標準？抑或是健康保險是否應給付？長遠來看，自動化科技輔助下的工作標準提高了，工作程序也改變了，連帶促使整體科技系統的調整，需要評估其利弊。對人們最直接的威脅將是工作機會遭 AI 機器人取代，但對各類職業或群體的影響不一。多數研究認為對年輕人、女性或新移民的影響程度較高，而各類職業的失業風險不同也可能是性別分工的差異所致。換言之，我們應該進一步思考，是否應建立某些政策主張，以確保全民都可共享 AI 科技發展的果實，而非僅是少數人獲利，而有多人得面對就業安全的威脅。

三、AI 的再現

社會人文研究者也值得關切 AI 的再現，是否複製偏見，還是創造新意。設計 AI 機器人時，有時需要設定生理性別，已有學者稱此是建立文化生殖器（cultural genital），藉此考察其性別邏輯。例如，強調服務的機器人是否多為女性，而需要建立權威感的，則以男性為主。已有一些研究發現，數位助理多設定為女性，而電影中的威猛機器人則為男性。聯合國科教文組織近期就發布報告，提出 AI 助理多為女性的性別刻板印象，需要改正。丹麥研發的 Q，就以建立無性別的聲音著稱，有利於改革現今語音助理的女性化現象。聯合國的報告也提出，改革 AI 的性別偏見，也需要側重鼓勵女性或少數族裔人才參與科技研發。

最後，吳嘉苓談到科技部近期提出的「人工智慧科研發展指引」，內容方向也呼應了本次論壇。然而，指引中像是「共榮共利」、「公平性與非歧視」、「問責與溝通」等等重要準則，在 AI 實際的操作上要如何運作，還需要許多實作的落實與交流。女性主義社群在「性別化創新」提出的性別研究方法，從如何決定主題、運用哪些概念及研究方法、在應用的時候要注意哪些社會環節，很值得作為重要資源。這有待與會者持續的創意與行動！

引言後的專題討論分為兩個部分，各邀請三位講者，分別以「AI——性別如何相互重組」及「AI 渴望性別的一百種方法」兩個主軸進行探討。

圖二：AI與性別論壇海報

Section I：AI——性別如何相互重組

主持人：林文源（國立清華大學通識教育中心教授）
主講人：吳秀瑾（國立中正大學哲學系教授）
　　方念萱（國立政治大學傳播學院副教授）
　　余貞誼（高雄醫學大學性別研究所助理教授）

聚焦在「AI——性別如何相互重組」，首先，國立中正大學哲學系暨研究所吳秀瑾教授採用數學博士 Cathy O'Neil 的演算法設計者視角切入，提醒我們由邏輯與數學運算公式組合成的 AI 機器學習演算法，其實不如我們想像的公正。演算法不僅反映設計者的價值選擇或偏見，而且會透過運算迴圈逐漸放大，但我們卻誤信 AI 會比一般人的判斷更為客觀公正。因此，在銀行信用評定的情境中，由於我們對特定職業或身分類別的群體存有偏見，則此偏見會影響銀行對於此群體的信用評等，且銀行信用評等結果甚至進一步影響我們對於該群體之知識與能力的判斷。

至於如何讓 AI——性別相互重組？吳教授觀察到國內社會文化中對性別的刻板印象具有高度相似性，致使科技產業的女性從業人口比例偏低，建議或可由提升女性工程師或科學家的比例著手。具體作法是進行高中數學教材與教法的翻修，把簡潔生澀的公式定理以平民化的方式表述，類似數學的白話文運動。當女性普遍具有足夠的數學能力，可以稽核 AI 機器學習演算法，才能消弭可能隱身在演算法背後的性別偏見。

其次，國立政治大學傳播學院方念萱副教授談的是大眾媒介裡反覆出現的女性機器人[1]遐想症（fembot fantasy），在電影裡彷彿成為一個新天地。但近年在大眾文化、學界中的討論，部分是與加拿大多倫多恐攻事件有關（Toronto van attack），而其中的攻擊者自稱是 incel（非自願獨身者），並在網路中發起了非自願獨身者的革命（incel revolution），且此現象已引起了西方社會廣泛地討論。經濟學教授 Robin Hanson 便認為，性機會較少的人所受到的對待與收入低下的人的情形相當，為了要求性的重分配（對比財富重分配），發出隱性的暴力威脅似乎是可以理解的。但很多人不同意這樣的觀點，如紐約時報的評論者 Ross Douthat 評論〈性的再分配〉一文時就引述了牛津大學哲學系 Amia Srinivasan 教授的看法，認為人有權去希望得到他們想要的東西，但性不是應得權利，成為

[1] 方念萱副教授解釋：女性機械人是一個能被程序化控制、以創造者的喜好當作行為標準的女人。

慾望的對象不是每個人的權利，但是什麼人是慾望的對象，什麼人不是，則是屬於政治問題。

方教授說明當代性的重分配的影響時，引述 Ross Douthat 所提供的以下三層面，幫助我們再思 AI 與性，AI 與性別的關係。一、性革命創造了新的贏家和輸家，以新的方式讓顏質高、金錢無虞、擅長社交的人獲得特權，並讓其他人陷入新的孤獨和沮喪，如同和其他形式的新自由主義去管制一樣；二、在新的經濟和技術變革中，兩性日漸無法理解對方，其間的社會和政治鴻溝加深，導致婚姻、家庭和性活動都在減少中；三、這反過來鼓勵人們——尤其是在現代性之下——把逃離一場革命的代價的希望，寄托在另一場還沒到來的革命上，無論它是政治、社會還是技術的革命。對很多顯然已經被進步拋在了後面的人來說，它即使不能帶來承諾的烏托邦，至少也能提供某種形式的補償。方教授總結相關文化評論的觀點後指出，在文化上，對於機器人的感情感受逐漸升高，終至無可抵禦，形成一種文化投降的樣態。且在當代文本、言說中看到的就是期待，帶著焦慮的期待，形成期待的政權（regimes of anticipation）。但 fembots 能否提供光明的性的未來？其實並非直觀而得，消費者必須繞道經過像是人造意識／機械姬裡的 Ava 這樣的經驗，才終能想像有著女性機器人陪伴的性的未來。

再次，聚焦在「我們希望演算法做對什麼？」，高雄醫學大學性別研究所余貞誼助理教授剖析演算法的運作機制（Algorithm ＝ Logic ＋ Control），試圖看見性別的痕跡如何滲入 AI 的運作。演算法所能認知到的事物本體，是一種後設資料的形式，並以數學的普世性為基礎，訴諸機械客觀性（mechanical objectivity）來找出肉眼未見、且不含藏人為偏見的模式和結構，以達成一種「客觀」、「嚴謹」的知識型態。然而，演算法不僅具有普世性的「邏輯」層面，也有可慎入人為偏見的「控制」層面。因此，對於演算法的探究，事實上是一種科技—文化政治（techno-cultural politics），由於其所運作的不僅是客觀的符碼物件，還包括各種依循演算目的而生的控制手段，因此若要解析其所呈現的世界如何而來，必須透過批判性的技術實作立場，從其過程、行動者、功能、和它的力量及其模式的侷限，來探究其運作的樣貌，並藉由理解其促成的影響的程度，來重新思考它的展望與方向。

余教授從四個面向來討論控制的環節是如何滲入主觀的性別意識形態，包括測量指標的選擇、資料集的選用、資料清理過程及數據詮釋。在測量指標的選定上，若以 AI 面試官為例，倘若好人才的評定包含情緒、個性、參與度、領導力和社交溝通能力等構面，這些構面的選取，以及以什麼特質作為這些構面

的評量標準，其實就蘊含了特定的價值關聯。在資料集的選用上，若我們「餵」給演算法的訓練資料本身即具有特定性質／傾向，演算法也會產出特定性質／傾向的結果。而在資料清理的步驟中，什麼資料要被留下／刪除的判準，也會影響最後的演算結果。最後，在資料詮釋時，雖然大家都說數據會自己說話、有圖有真相，但不論數據或圖，都仍需仰賴詮釋的工作，才能將內容帶回數據中，深化其成為厚數據的可能性。

透過剖析演算法的控制過程，余教授認為我們需要進一步問的是，我們希望演算法做「對」什麼？是藉其去反映一個既存的社會現象，還是一個有著「更大的善」的形象企圖？這樣的價值關聯議題，正顯示了，演算法所涉及的不僅是純粹技術問題，而含有應該審慎思考的經濟、政治和文化議程。所以積極面對演算法的態度，是走入演算運作的流程，去看見其中的各種資訊規模、流向、處理和決策過程，讓各種價值關聯的開放討論，成為可能。

Section II：AI渴望性別的一百種方法

主持人：楊谷洋（國立交通大學電機工程學系教授）
主講人：謝舒凱（國立臺灣大學語言學研究所副教授）
　　黃從仁（國立臺灣大學心理學系助理教授）
　　陳宜欣（國立清華大學資訊工程學系副教授）

在「AI 渴望性別的一百種方法」討論主軸中，首先是國立臺灣大學語言所謝舒凱副教授從語言的角度看性別偏見、自然語言處理怎麼處理偏見，以及談談他目前看到的可能問題跟想法。謝教授認為目前所談的 AI 偏見，所有都是來自人類的偏見。為了探討性別意識形態如何反映在語言使用，性別偏見與對立立場如何透過語言來形塑論辯，謝教授選擇了一個專門做同志文化（LGBT）的社會語言學分支，進行一個簡單的分析，發現代名詞「我」、「我們」在同志族群的文本中較異性戀文本更明顯的被大量使用。而社會心理學的討論認為，若文本中大量使用第一人稱代名詞時，可能反映作者在情緒上的焦慮感與絕望，或是呈現出作者身處於社經地位較低的狀態。另外，語詞的使用（同性戀／同志）反映了不同的立場，或是不同詞語結構（非常規名物化成「同性戀行為」）乃是反映某種意識形態。

然而，謝教授認為，比較令人擔心的其實是現在的 AI 模型可以通用（相同模型可以用在語言處理、圖像辨識或其他 AI 應用），AI 通用模型可自行產生假的照片、假論壇帳號，並且發表假的評論，造成虛實之間的撲塑迷離，真假難辨。若要解決自然語言處理可能產生的偏見，除了選擇較不具偏見、較平衡，或刻意製造沒有偏見的語料讓機器學習；在模型端則運用技術刻意懲罰具偏見的詞語，甚至是讓模型透明化並具有可解釋性，這樣我們才可以由人文社會的角度去修正，但這些方法並沒有真正解決偏見的問題。比較核心的議題應該是標籤與連結的計算本質是效能優化而不是價值多元，另外就是監督式學習需要標準答案，但是一個理想的、沒有偏見的社會應該是沒有標準答案的。所以謝教授呼籲在發展技術時，應把社會責任當成首要目標，另外就是 AI 也需要多元發展，讓不為商業服務的 AI 成為可能。

繼之，國立臺灣大學心理學系黃從仁助理教授以「AI 如何物化人類？人類渴望變漂亮的 100 種方法」為題探討機器學習的潛在偏見，並嘗試理解人類與 AI 評審如何審美。黃教授利用一個具有 500 張亞洲女性臉部照片及其顏值的公開資料庫，訓練一個卷積神經網絡（Convolutional Neural Network, CNN）來作為模仿人類評審的機器審美官。此機器評審在完成訓練後進行測試，其評分與真人評審的評分相關性可高達 r ＝ 0.7-0.8，顯示此機器與人類的審美觀接近但並非完全相同。接著，黃教授使用一系列的「可解釋性 AI」（Explainable AI）的方法來瞭解機器評審的決策標準和傾向，示範如何打開機器決策的黑盒子。

黃教授將機器評審視為像人的研究受試者，分別使用觀察法與實驗法的方式去瞭解其決策傾向。透過觀察法比較高分組與低分組的平均臉形後，可得知機器對於臉較瘦長、眼睛較大、鼻子較長、嘴角上揚的照片評分較高。此外，透過檢視機器將注意力擺在哪些像素（pixels）上做出決策，我們可觀察到機器會因為深眼袋和大鼻孔而給出較低的評分。而透過類似於心理物理學（psychophysics）的實驗法，研究者則可以根據照片中的某種物理特性對照片做分類或排序，系統化地檢驗機器評審是否會因為這個物理特性變化而有評分的高低差異。我們都知道對稱性是人類評定美醜的標準之一，用此方法檢驗後也發現機器評審有將此因素納入顏值評判中。另外大家都說一白遮三醜，但卻沒有科學實證的根據。而透過此方法則可發現機器評審偏好打光較好的「亮白臉」（即在色彩空間 HSV 中有較高 Value 值的臉），但卻沒有偏好臉上較無血色的「慘白臉」（即在色彩空間 HSV 中有較低 Saturation 值的臉）。因此，黃教授以自己的照片做了一個有趣的實驗：他先以一側臉進行鏡像化，使左右邊臉完全對稱，另外再把照片中的皮膚塗白似麥當勞叔叔，最後讓機器評審評定顏值。結果顯

示，這些影像操弄的步驟可以愚弄 AI 審美官，大幅提升機器評斷的顏值分數。黃教授表示：我們常畏懼成為 AI 決策下的受害者，但透過如上的可解釋性 AI 方法，其實我們能「役 AI 而不役於 AI」。

最後，國立清華大學資訊工程學系陳宜欣副教授則是探討如何將性別因素納入，幫助 AI 機器學習。陳教授曾參與「科技部性別與科技相關規劃推動計畫」，為該計畫建置各學科學生的性別比例等等的視覺圖表，希望讓各政府部門的性別統計資料更容易使用。另外也協助「女科技人電子報」的編輯出版，不久前剛發表一篇文章，利用國內網路資料的文字探勘，探討人們使用文字時的性別偏見。

科技的「性別化創新」是國際上一個重要的新興領域，但實際進行時，有時會面對原始資料難以判斷性別的窘況，因而難以繼續研究。那是否可以用 AI 來猜資料的性別？陳教授和團隊蒐集了許多不同年代的名字，發展猜名字性別的 AI。分析結果發現，名字是有時代感的，而且名字第二個字的部首對於性別判定最為重要。如果演算法進一步把女生名字的用字群，扣掉男生名字的用字群，剩下的字（只有女性使用的字）大多與描述美貌身材有關；若將男生名字的用字群扣掉女生名字的用字群（得到只有男生使用的字），大多是用來描述一些外在的事物，例如：山、豬頭、棋子、臣、山寨、種植、兵器等，這與我們直覺印象不同。而以 AI 判斷名字所屬性別的正確率達 9 成以上，遠高於人力判斷的 6 成左右。

此外，陳教授思考，如果在分析情緒與精神疾病時，關注性別因素並加以區分，是不是有不同的發現？過去一些研究結果認為，女性在談話時較常表達悲傷或焦慮，而且較常提及家人或社群；男性常出現生氣或咒罵的字眼，且少用第二、三人稱。從陳教授的研究結果發現，如果在訓練 AI 時加入了對於男性與女性的情緒偵測器後，AI 的情緒分析正確率都提高了。此外，如果把男性與女性躁鬱症的例子分開來看，AI 正確分析女性躁鬱症的比率比分析男性的還高。也就是說，在某些特定的小眾案例，先將資料以性別來區分，較少量的資料反而能訓練出較高正確率的神經網絡。因此，在擔心 AI 歧視性別的另一面向，或可換個角度來思考，不需要刻意弭平資料的性別差異，而是加以正視，使用這個差異來達成對人類社會的瞭解。