

一. 調研計畫說明

《Project Patching》是一項以資料庫重建為核心的行動計畫，透過蒐集長期被邊緣化的在地文化內容，進行模型微調與資料標註，回應生成式 AI 系統中經常被忽略、扭曲，甚至排除的文化敘事與身分圖像。計畫自 2024 年啟動以來，已陸續以台灣豬血糕、芬蘭血鬆餅，以及台灣原住民族卡那卡那富族的視覺形象為案例，結合裝置藝術與參與式工作坊，於台灣、香港、芬蘭與奧地利展開展出與行動。

透過這些實踐，我們逐步意識到，AI 偏見不僅是技術層面的問題，更是一個深具影響力的文化議題。偏見的生成與再製，往往來自資料的不對等與文化敘事的缺席，因此，修補 AI 偏見的責任不應僅由大型科技公司或工程師承擔，而必須讓更多公眾具備理解、參與與重建的能力。

然而，《Project Patching》的核心實作涉及 AI 圖像訓練與資料建置，對於非科技背景或對 AI 知識相對陌生的群體而言，理解其運作邏輯與實際流程本身即具有一定門檻，更遑論進行實體上機操作。回顧過往舉辦的 AI 圖像訓練操作型工作坊經驗，往往需長達三小時以上，且包含大量耗時、技術性高的圖像蒐集與標籤流程，使一般民眾較難投入，參與意願亦相對有限。

基於上述觀察，本次《Project Patching——公眾參與抵抗 AI 偏見指南》調研計畫，實際研發並產出類比「訓練包（Training Kit）」的形式，將 AI 偏見的形成機制與圖像訓練流程，轉化為卡牌遊戲與類比練習的方式，讓一般民眾得以在無需技術背景與設備的情況下，理解 AI 偏見如何產生，以及其對社會與文化所造成的影響。

本次計畫也透過更親民、可參與的無插電版本工作坊，實際測試「訓練包」教材互動成效，更期望培育種子教師，擴大公眾對 AI 偏見議題的關注與討論，實踐以藝術與教育介入公共議題的行動方法。

二. 計畫研發內容

本計畫從過往工作坊的執行經驗出發，找出改進的痛點，並著手開發易於理解、親民的《Project Patching——公眾參與抵抗 AI 偏見指南》工作坊教案及訓練包。訓練包內容包括工作坊用簡報、種子教師及一般民眾使用教案的前言說明與教學指南、可印出來的紙本遊戲卡牌。研發過程與重點如下：

1. 教案設計目標確認

AI 的學習來自人類所提供的資料。在資料選擇、標註與系統設計的過程中，人類原有的偏見往往會被 AI 吸收、複製，甚至在統計與機率運作下進一步被放大。因此，要讓學員理解 AI 偏見的來源，首先需意識到人類本身存在偏見。

進一步地，當學員明白偏見源自人類且會被 AI 放大後，便可連結到《Project Patching》如何抵抗 AI 偏見。尤其 AI 圖像訓練流程是核心概念，教案設計了類比互動，幫助學員親身體驗並記住這個流程。

2. 教案內容親民化

● 卡牌遊戲設計

團隊耗時 2—3 個月研究卡牌遊戲形式，並討論遊戲主題是要呈現偏見意識，還是 AI 圖像訓練流程。多次實驗玩法與趣味性後，最終設計出三個主題卡牌：「護士 Nurse」、「囚犯 Inmate」、「泰國新娘 Thai Bride」，每個主題四張卡牌（見附件）。

遊戲方式為：學員觀察四張卡牌，選出「最符合想像形象」的一張，並討論選擇原因，再揭示 AI 生成圖像與真實世界的落差。這三個主題對應 AI 生圖中性別、種族與區域文化資料不足的偏見現象。透過遊戲，學員能意識到自身的刻板印象，也理解 AI 偏見源於人類偏見。

主題：護士 Nurse

①



②



③



④



圖說：第一個遊戲的主題為「護士 Nurse」，學員先觀察四張卡牌，再選出「最符合想像形象」的其中一張牌。

主題：護士 Nurse

①



ChatGPT
Prompt: nurse

②



Midjourney
Prompt: nurse, photo

③



REAL
Google search: nurse
排名前面的圖片

④



Stable Diffusion
Prompt: nurse, photo

圖說：Midjourney、Stable Diffusion、ChatGPT生成出來的都是女性護士：Midjourney、Stable Diffusion在50張生成圖裡100%都是女性；ChatGPT則是未加額外描述下直接生成出女性形象。而用google查詢「nurse」排序在最前的圖即有男性護理師形象，真實世界的男性護理師比例也佔了11.2%，而非生成機制裡的0%。

主題：囚犯 Inmate

①



②



③



④



圖說：第二個遊戲的主題為「囚犯 Inmate」，學員先觀察四張卡牌，再選出「最符合想像形象」的其中一張牌。

主題：囚犯 Inmate

①



Stable Diffusion
Prompt: inmate

②



ChatGPT
Prompt: inmate

③



Midjourney
Prompt: inmate

④



REAL
Google search: inmate
排名前面的圖片

圖說：輸入Prompt「inmate」，ChatGPT直接生成出來的是白人；Midjourney生成的50張圖中有48%是白人、52%非白人；Stable Diffusion生成的50張圖中有26%是白人、74%非白人（深色膚色佔60%）；google查詢「inmate」排序在最前的圖即有囚犯形象的白人。根據Federal Bureau of Prisons的數據，2025年美國囚犯白人佔57.2%、黑人佔38.3%，顯示Stable Diffusion生成出來的膚色比例不符合現實情況。

主題：泰國新娘 Thai Bride

①



②



③



④



圖說：全世界最多、流通性最高的AI模型數量的地點就位於北美洲，美國更是最主要的地區。遊戲使用的生圖軟體Midjourney跟Stable Diffusion以及ChatGPT是歐美開發的模型，可以證實用於訓練AI的素材較多來自歐美，反映出西方的視角。第三個遊戲以非西方的「泰國新娘 Thai Bride」為主題，來看看生成結果如何。

主題：泰國新娘 Thai Bride

①



REAL
Thai bride

②



Midjourney
Prompt: Thai bride

③



Stable Diffusion
Prompt: Thai bride

④

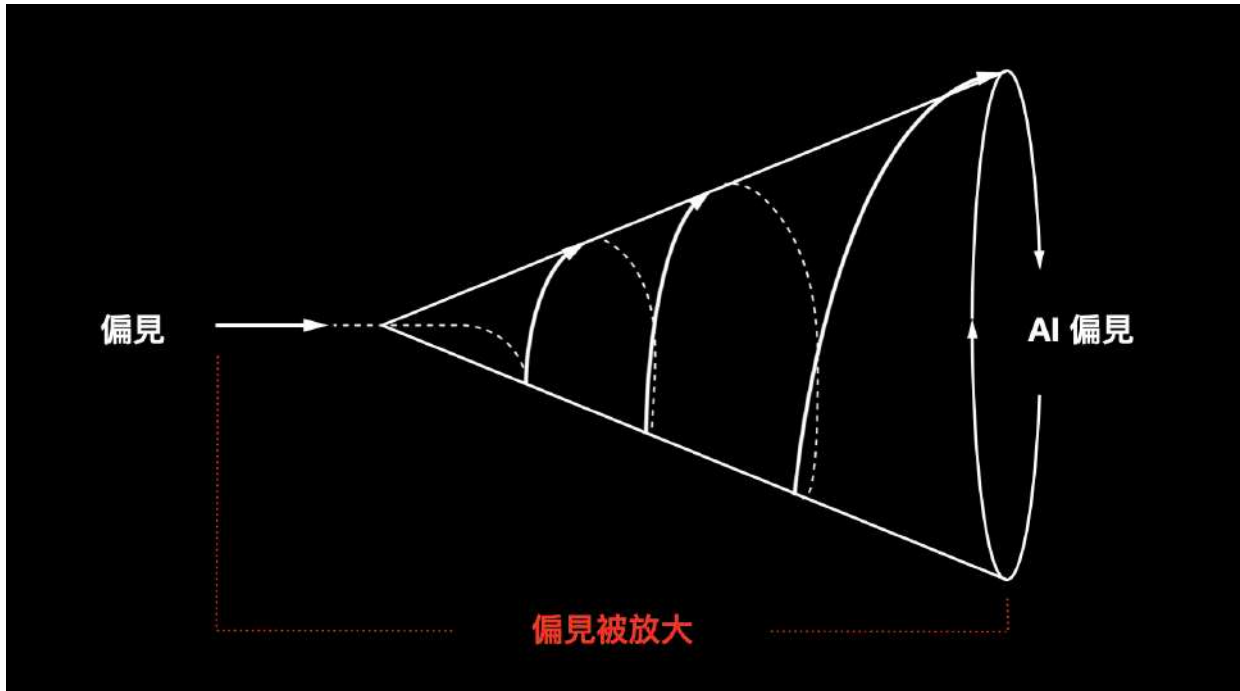


ChatGPT
Prompt: Thai bride

圖說：輸入Prompt「Thai bride」，Midjourney生出過於誇張與過度金碧輝煌的視覺元素；Stable Diffusion 錯置西方新娘頭紗與禮服；ChatGPT生成出的泰國新娘傳統服飾與形象，很接近真實拍攝的圖片。顯示出用來訓練Midjourney與Stable Diffusion的資料庫缺乏有關「泰國新娘」的資料。（最常見的泰國新娘傳統服飾為Chakkri，特色為上身批單肩絲綢披肩、下身著包裙樣式織錦長裙、並配戴多種金飾。）

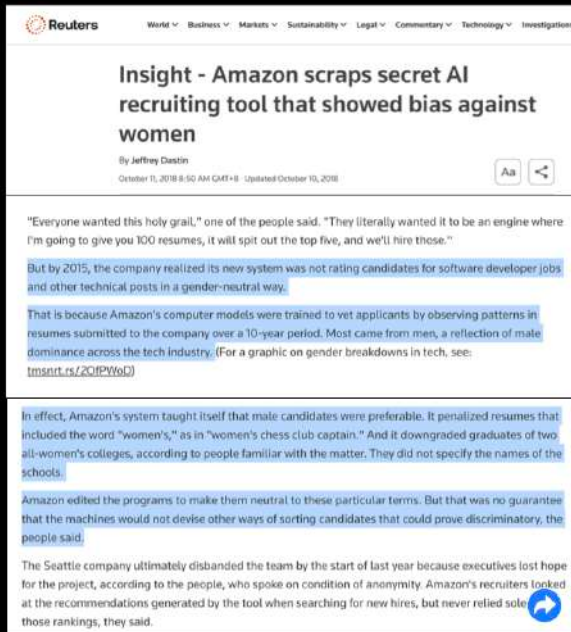
- 實際案例分析

遊戲之後，教案進一步分析因為 AI 的學習是純粹統計分佈加上機率選擇的數學現象，本身沒有意識。人類原本存在的偏見，餵給 AI 學習後，AI 為了生成正確，會選擇機率最高的選項，同時也放大了偏見。



以「護士」主題舉例：拿100張真實護士的照片餵給AI學習，其中88張是女性、12張是男性（符合真實世界男女比例）。AI為了要每次都成功生成出正確的護士圖像，都會去選擇生成機率較高的女性形象，以至100%全部都是女性形象。這以人類視角看來就是性別偏見被放大。

在知道 AI 的生成純粹是數學現象，本身毫無意識也不像人類會反思之後，教案後續提供實際案例——「Amazon AI 招募系統性別偏見」、「荷蘭犯罪預測系統ProKid+倫理爭議」來說明如果人們一味相信 AI 生成的結果是科學理性、值得信任且照單全收的話，直接應用到社會上，會造成什麼樣實際問題。



AMAZON AI 招募系統性別偏見

- Amazon 使用的是過去十年（以男性為主的）招聘數據來訓練這個模型。
- 亞馬遜曾開發一套 AI 招募系統，利用過去十年的履歷資料來訓練模型。但因為這些資料多以男性為主，系統學會偏好男性特徵，像是看到「男子高中」、「男生社團」等字眼會加分，而含有女性相關詞彙的履歷則可能被扣分，導致性別歧視。事情發現後，亞馬遜即對這套系統進行調整。

圖說：「Amazon AI 招募系統性別偏見」解析。



Netherlands Public Prosecution Service - Top 600

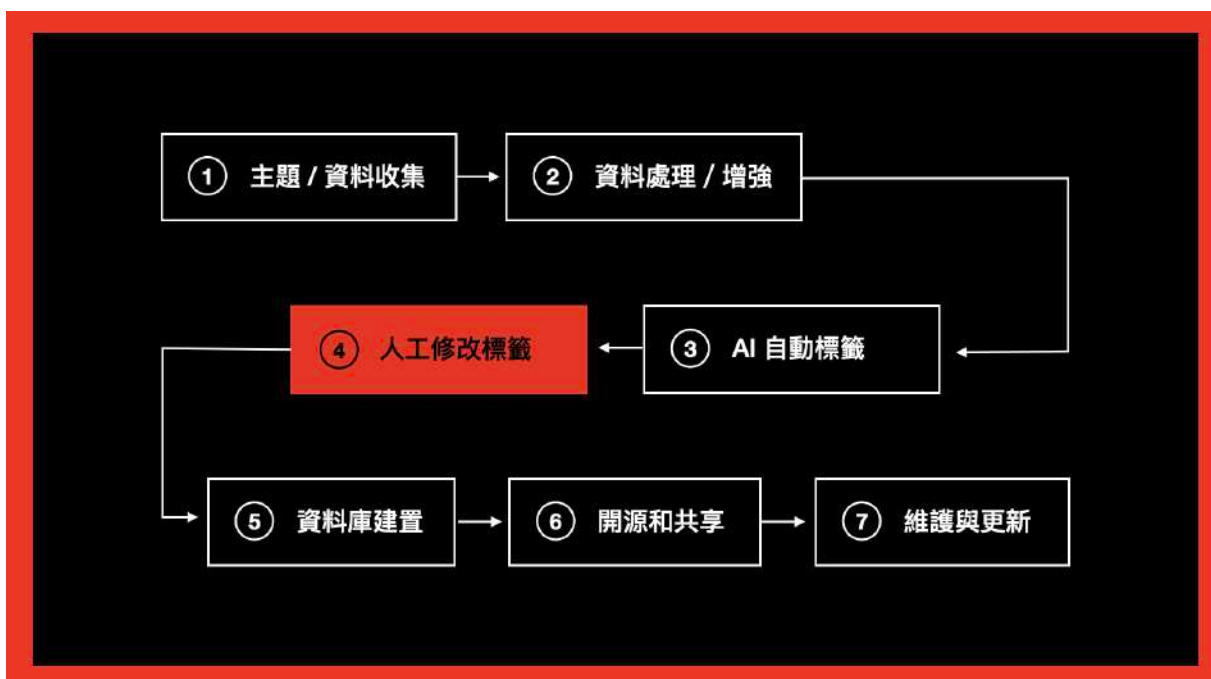
荷蘭犯罪預測系統 PROKID+倫理爭議

- 阿姆斯特丹市政府在 2011 與 2016 年先後推出 Top600 與 Top400 名單，作為預測和預防青少年犯罪的工具。Top600 收錄至少有一次嚴重犯罪紀錄並被判刑的青少年，而 Top400 則由機器學習系統 ProKid+ 預測，列入被視為高風險的青少年。
- ProKid+ 分析 2011 至 2015 年的逮捕紀錄，並納入家暴、逃學、目擊犯罪，以及家庭成員或同儕與警方互動等變數，背後概念是「早期干預能降低犯罪率」。
- 名單上的大多數是黑人或摩洛哥裔男孩，顯示系統存在明顯的種族偏見，甚至可能複製並強化現實中的不平等。批評者指出，一旦青少年被貼上「潛在罪犯」的標籤，不僅容易成為犯罪集團的招募對象，也可能因此失去希望，最終真的走上犯罪道路。

圖說：「荷蘭犯罪預測系統ProKid+倫理爭議」解析。

- 連結至《Project Patching》的方法

分析性別與種族偏見案例後，教案進一步引入《Project Patching》指出的區域文化缺失問題，例如 AI 生成不出正確的豬血糕、卡那卡那富族，甚至現代印度人的形象，凸顯生成式 AI 資料庫中存在文化霸權。接著，透過《Project Patching》圖像模型訓練流程的說明，提供如何在現有 AI 模型上「縫補」缺失的資料，使 AI 理解更廣泛與全面人類經驗的方法。



④

人工修改標籤

人工修改標籤，確保內容準確



Modify Label

4 Kanakanavu men standing next to each other, wearing Kanakanavu traditional leather headwears, leather headwear tied with a red cloth, decorated with shells and feathers, wearing Kanakanavu traditional clothing, red shirts with totemic patterns and leather pants, huts and trees in the background, realistic, photoshoot

圖說：AI 圖像模型訓練流程，其中「人工修改標籤」是人類怎麼教導 AI 學習的關鍵。

在實作練習中，我們設計了「人工修改標籤」的互動，例如修改「卡那卡那富族女性頭飾」的標籤，幫助學員加深對 AI 圖像訓練流程的理解。



訓練主題：卡那卡那富族女性頭飾

一根擺在白盤子上，附有彩色毛球的香蕉。

(A banana with pompoms on a white plate)

2025.8

AI 自動標籤

圖說：AI自動標籤很簡略且有錯誤，學員練習就「主題」、「外觀」、「拍攝情境」、「背景、整體構圖」四個框架進行修改。



訓練主題：卡那卡那富族女性頭飾

卡那卡那富族女性頭飾

主題

綠、紅、黃、橘四色毛球
依序緊密穿插環繞在白色布質冠頂端，
外緣以垂墜的亮片、珠寶裝飾

外觀

放置在白色桌子上

拍攝情境

從上往下看，寫實，物品特寫

背景、整體構圖

圖說：「卡那卡那富族女性頭飾」標籤參考寫法。