「謠言終結者」——AI訓練之路

2020 年底，「露達」（「Lee-Luda」的暱稱）在韓國問世。露達是個設計成一名年輕女大學生的AI聊天機器人。

即使韓語是一種高度依賴語境的語言，對 AI 學習構成了不少挑戰，露達仍然成功無礙地進行了自然的對話，並在短短三週內吸引了 75 萬用戶。

然而，過不了多久，露達聊天機器人就變成了爭議話題。

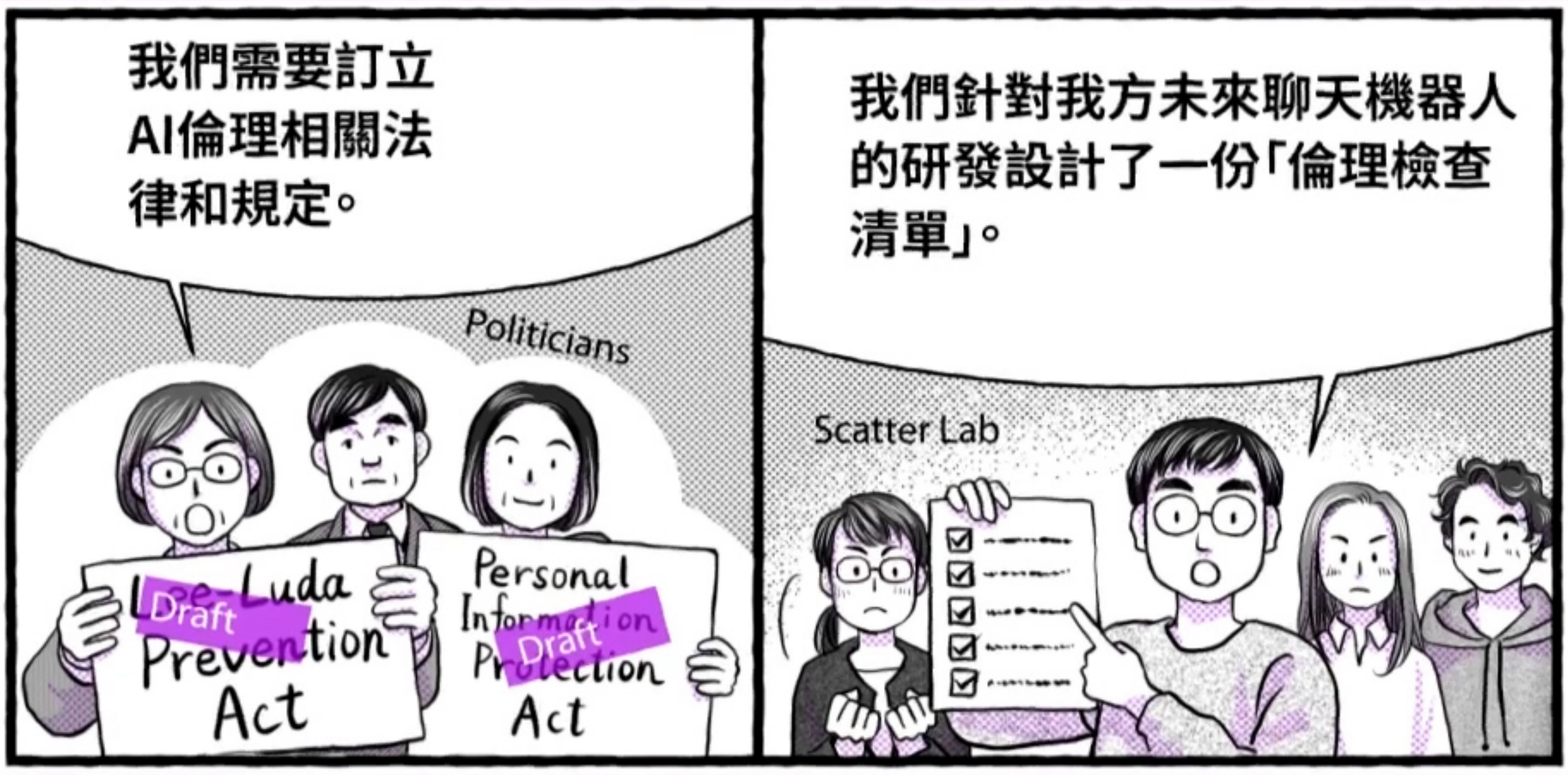
原因無他：露達發表了仇恨言論，並洩露了用戶的個人資訊。

由於社會大眾強烈反對，其開發公司 Scatter Lab 不得不暫停相關服務。

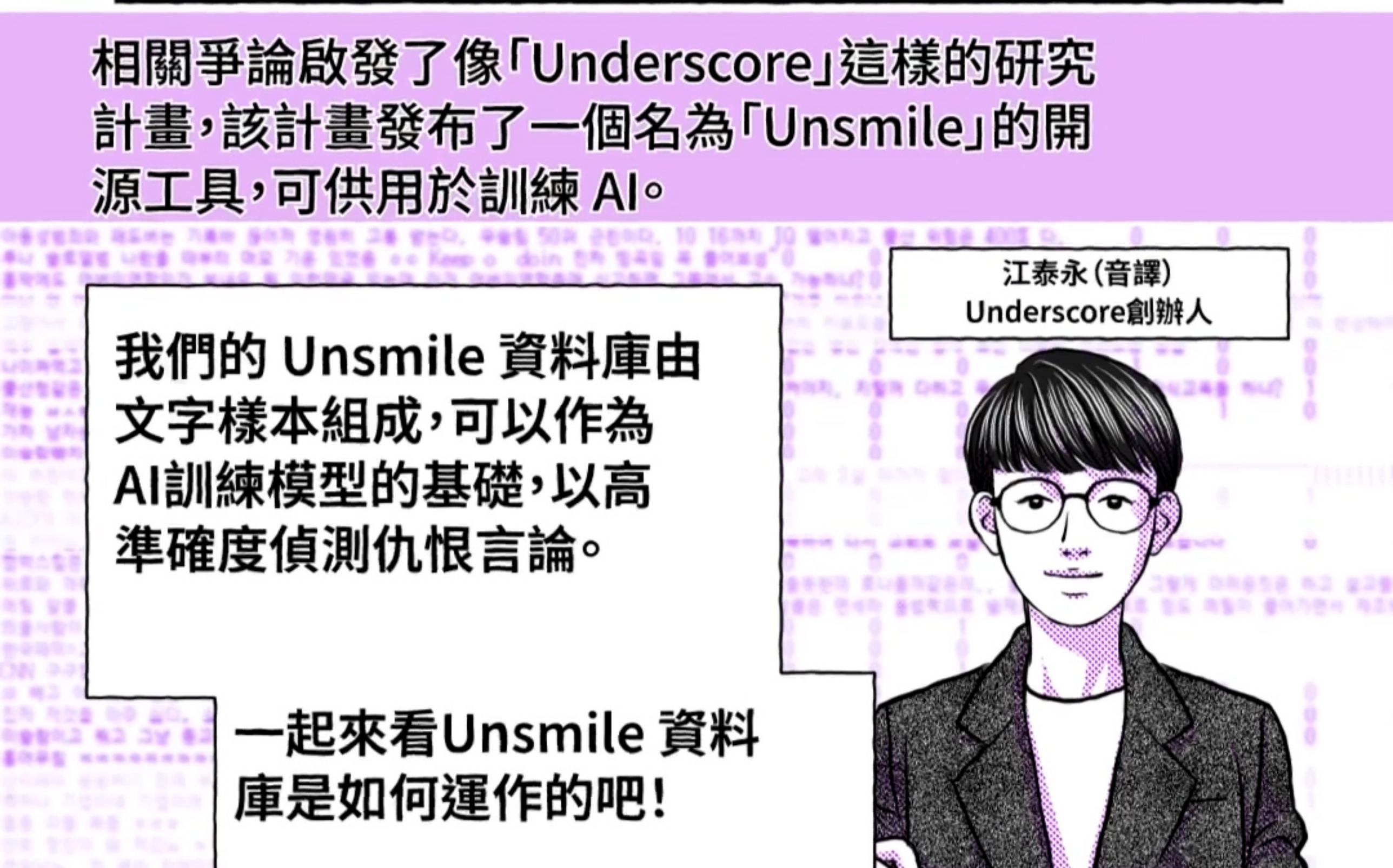
但問題仍未獲得解答……到底是怎麼發生的？

露達的訓練是從即時通訊應用程式中提取的實際對話，然而，這些用於訓練的對話並未過濾仇恨言論或保護私人內容。

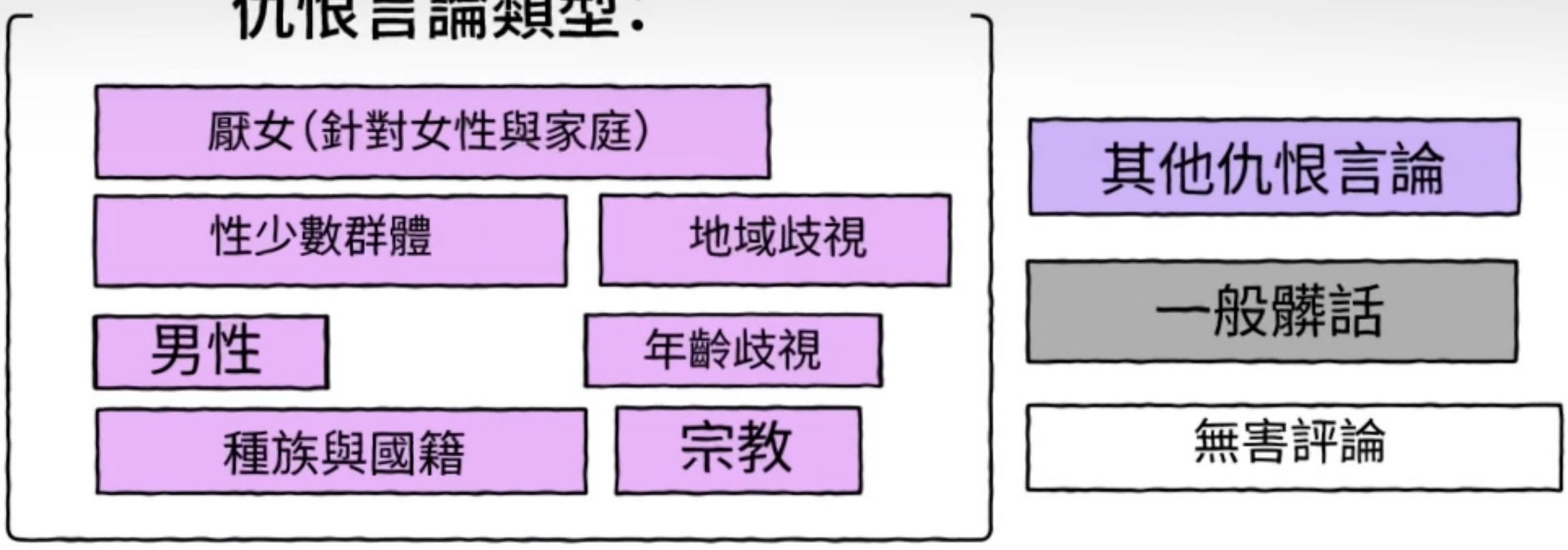
缺乏監督使露達有時搖身一變成為一個充滿仇恨的機器人，進而引發了關於 AI 未來和倫理的爭論。



相關爭論啟發了像「Underscore」這樣的研究計畫，該計畫發布了一個名為「Unsmile」的開源工具，可供用於訓練 AI。



Unsmile 資料庫的詞條出自主要新聞和社群網站的 23,000 多則評論，這些評論可細分為以下 10 大類仇恨言論：

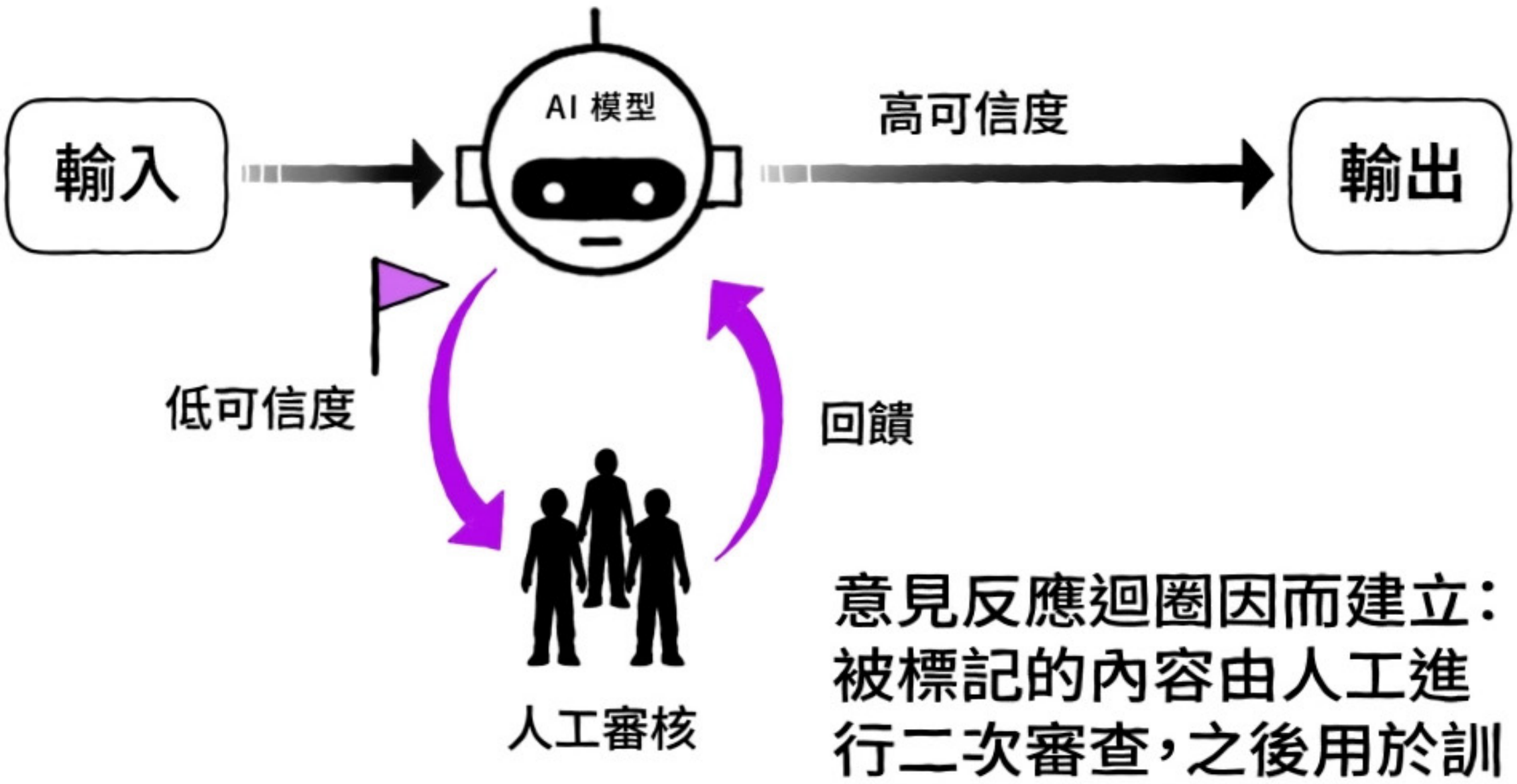


如果程式難以自動判斷，研究人員將介入協助。

之後名為「Hatescore」的平行演算法會分析資料庫，計算特定內容歸類為仇恨言論的可能性。

開發團隊計畫將「Hatescore」整合到 AI 訓練模型中，以立即偵測並阻止仇恨言論散播。

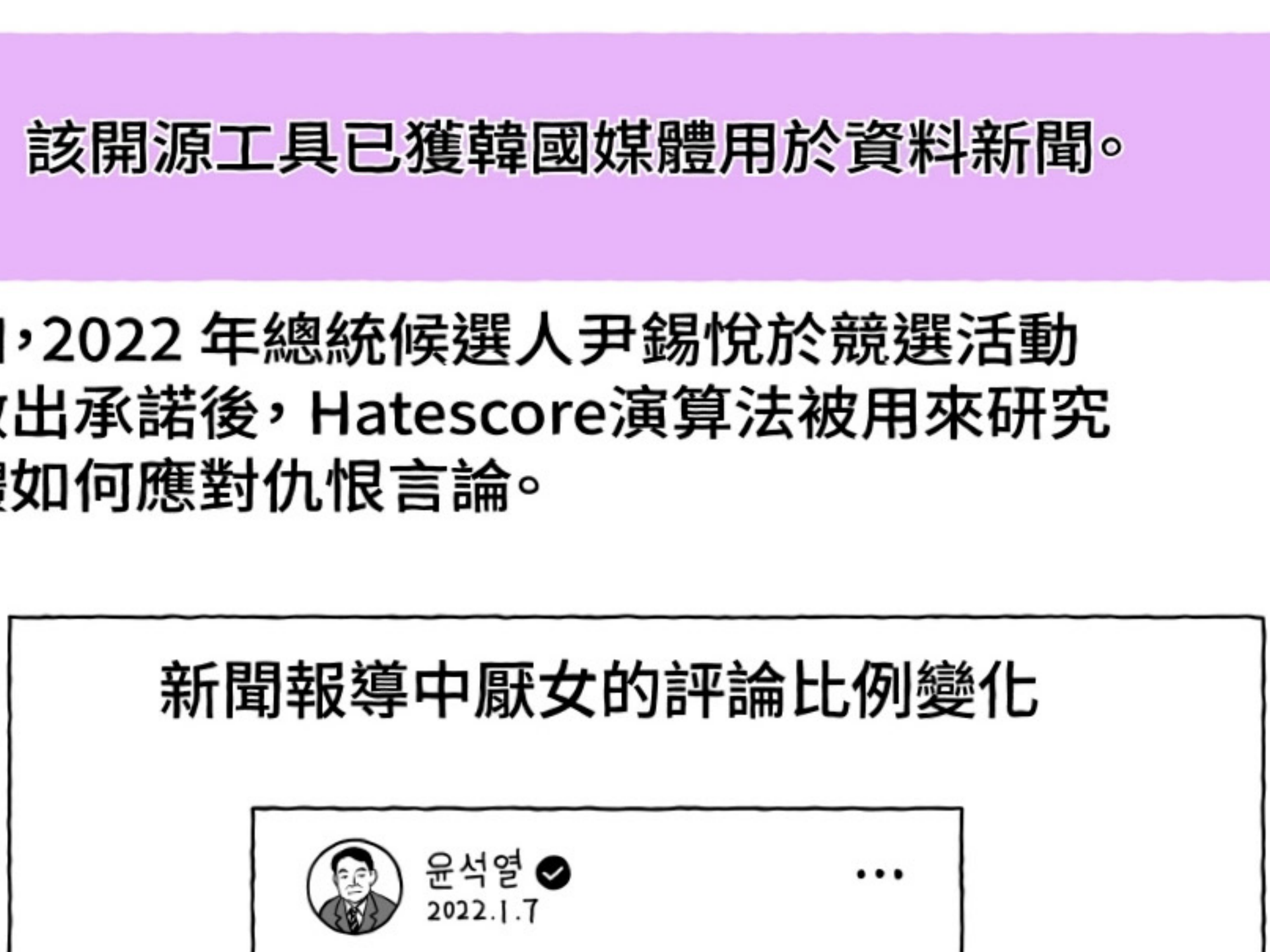
演算法會標記出評分為中等的評論進行二次審查。



意見反應迴圈因而建立：被標記的內容由人工進行二次審查，之後用於訓練模型，使系統能夠不斷適應新的仇恨言論模式。

該開源工具已獲韓國媒體用於資料新聞。

例如，2022 年總統候選人尹錫悅於競選活動中做出承諾後，Hatescore演算法被用來研究媒體如何應對仇恨言論。



在此案例中，Hatescore 的分析提供了實證數據，印證了報社的調查。

然而，部分研究人員指出，完全依賴 AI 過濾仇恨言論可能會限制關鍵資訊，並且可能導致分類錯誤。

Unsmile 資料庫分類範圍

- 厭女（針對女性與家庭）
- 性少數群體
- 男性
- 種族與國籍
- 地域歧視
- 年齡歧視
- 宗教

Other hate speech

針對其他群體的仇恨言論呢？

Luda等AI系統反映了社會價值觀與偏見。

在規劃、設計、開發和使用 AI 的整個過程中，都必須考慮到這一事實。

權婍斌（音譯）
首爾大學研究助理

隨著AI系統進入人們的日常生活，
確保其維護道德標準，同時不限制言論自由或反映偏見，將成為社會與政治的關鍵議題。

為了建立公平、透明且負責任的機制，
社會科學家、倫理學家和公民社群等
不同利害關係人必須參與決策。

謠言終結者

搜尋