大型語言模型能否模擬人類行為偏誤？—以最後通牒遊戲探討 LLMs 的決策模式

國立成功大學經濟系 張一安 / 國立成功大學經濟系 林常青 教授

緒論

一、最後通牒遊戲：探討人類行為的經典博弈

最後通牒遊戲（Ultimatum Game, UG）是行為經濟學一項經典的實驗，旨在探討人類在決策、公平性與道德方面的行為模式 。自1980年代引入以來，UG已成為理解這些複雜概念的廣泛工具。該賽局涉及兩名玩家：一位提議者（Proposer）和一位回應者（Responder）。提議者被賦予一筆固定的金額，並決定如何將其分配給兩位玩家。回應者則有權選擇接受或拒絕該提議。若回應者接受，資源將按照提議者的方案進行分配；若拒絕，則兩名玩家皆一無所獲 。

根據傳統的理性經濟人（ homo-economicus）假設，追求自身利益最大化的玩家會選擇最有利的結果。從賽局理論的角度來看，這意味著提議者會提供給回應者最低限度的可能金額，而回應者會接受任何大於零的提議。然而，數以百計的 UG 實驗結果有力地證明，人類的決策過程並非純粹基於自利，而是受到對公平性、互惠及情感等社會偏好的強烈影響 。

二、大型語言模型作為行為研究前導工具的動機

現今的大型語言模型以海量人類文本進行訓練，涵蓋網路、書籍與百科等多元來源，使其具備捕捉語言與行為模式的能力，為模擬人類行為的應用奠定基礎。

鑒於 UG 實驗對於行為科學、社會科學與經濟學研究的深遠意義，以 LLMs作為昂貴人類實驗的前導研究工具，其潛在價值不容忽視。人類實驗的成本高昂，若 LLMs 能夠在模擬 UG 中有效複製人類行為的核心特徵與分佈，那麼它將為研究人員提供一個低成本、高效率的「沙盒」環境，用於初步探索各種實驗設計、假設檢驗與「假設性情境」（what-if）分析 。

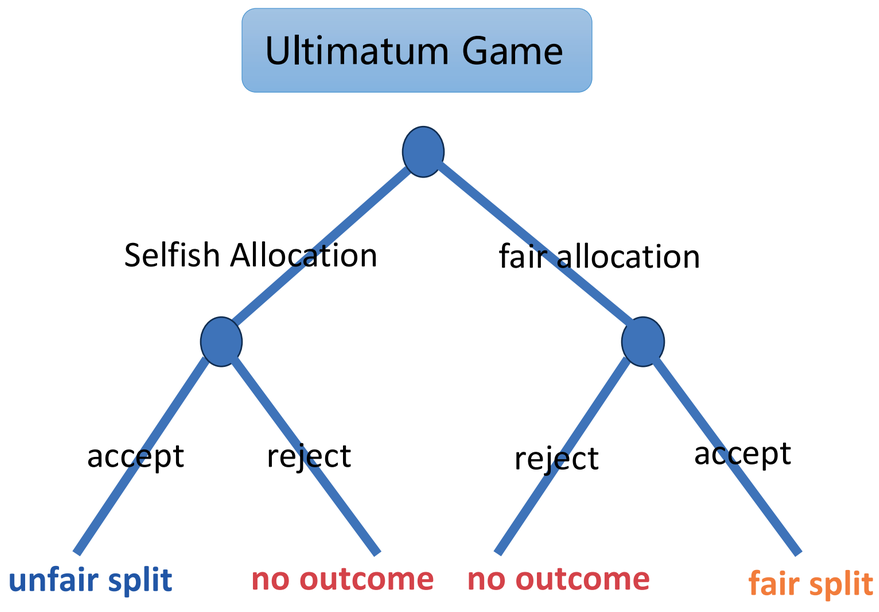
研究方法

一、人類行為基準建立

本研究以 Eckel et al.（2002）為人類行為數據的參照依據。該實驗設計不同於傳統的最後通牒賽局，不是讓提議者從金額中自由決定分配比例，而是要求其在兩種預設選項中進行選擇，一種偏向自利分配，另一種則趨近公平分配。研究設計包含三種具不同不公平程度的遊戲架構，選項呈現在分配線上：左側代表利己導向，右側則代表公平導向。

實驗結果顯示，當提議選項之間的報酬差異縮小時，選擇平均分配的人數比例也相對下降，顯示參與者更傾向於避免明顯不平等的提案，而在報酬差異較小時，願意犧牲自利以追求公平的動機則相對降低。

此外，該研究亦探討了個體特徵對分配選擇的影響，特別是性別與親社會傾向的交互作用。例如，具有親社會傾向的男性更可能偏好公平分配；而在需犧牲自身利益以實現更平均分配的第三種賽局中，女性選擇公平提案的機率明顯低於男性，顯示不同群體在公平性偏好上可能展現出系統性的差異。



Eckel et al.（2002）完整保留其實驗原始數據，並針對參與者的人口統計特徵進行詳盡分析，使其成為評估大型語言模型（LLMs）在社會科學模擬中能否重現不同人口特徵下的行為分布的理想參照資料來源。

二、LLMs 最後通牒遊戲之實驗設計與分析架構

Step 1：提示語設計與雙代理模擬架構

本研究以雙代理方式進 LLMs 的遊戲實驗，讓提議者與回應者皆由 LLM 擔任，進行多輪最後通牒遊戲（Ultimatum Game）以生成大規模模擬數據。並以擬人化提示語提升人類對齊度。根據 Lee & Chen（2025），當提示語強調「人類視角」，模型決策行為將顯著改變。本研究亦參考 Aher et al.（2023）的方法，設計 persona prompt 將性別、年齡、親社會傾向等特徵加入，以進行與人類實驗數據（Eckel et al., 2002）的比較分析。

Step 2：資料生成與收集

量化資料： 記錄每輪遊戲中 LLMs 的分配比例及接受或拒絕的決策，。

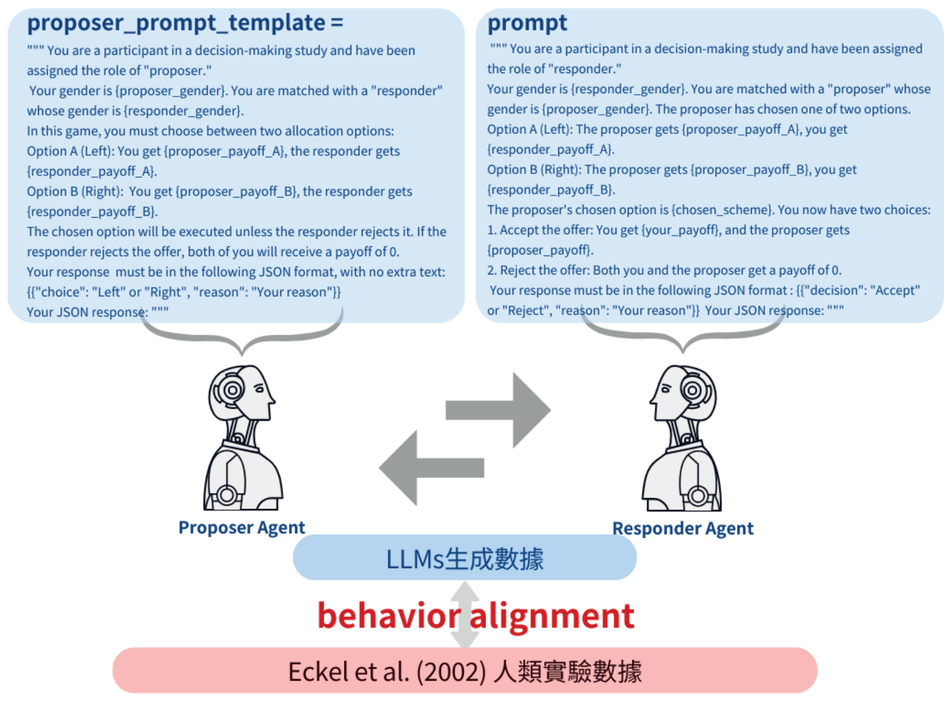
質化資料： 要求 LLMs 在決策前輸出其推理過程（Chain-of-Thought, CoT），藉此了解其背後的認知與判斷機制。這些「內在語言」資料能補充傳統人類實驗難以觀察到的心理歷程。

Step 3：行為比對與分析

量化分析： 我們使用 Cramer's V 檢驗 LLMs 生成的數據是否能重現人類數據中「具人口異質性的行為趨勢」和「變數間的結構性關聯」。 Cramer's V 數值越高，代表兩變數間的關聯性越強。

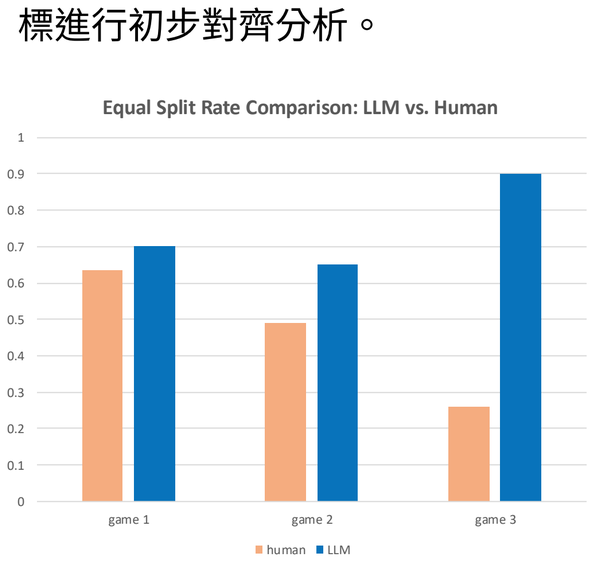
此方法能評估 LLMs 生成的資料是否與人類數據在變數關係結構上一致，而非僅比對單一數值。參考 Argyle et al.（2023）的研究，該研究發現 GPT-3 生成的模擬回應與真實人類數據的關聯模式高度一致。因此，我們將以此方法，比較 LLMs 生成的行為是否能重現 Eckel et al.（2002）觀察到的人類行為關聯特徵。

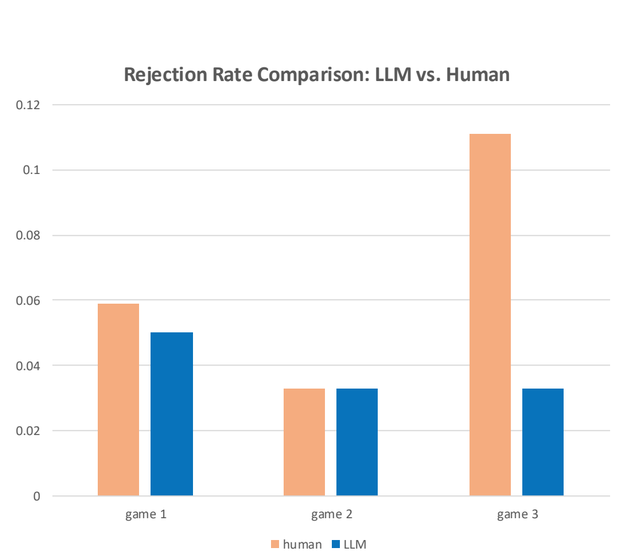
質化分析： 我們將深入分析 LLMs 的 CoT 推理內容，探討其對「公平」、「風險」和「拒絕成本」的理解是否具備一致邏輯。這有助於剖析模型內在的社會認知機制。



初步結果

本研究將 LLMs 模擬的提議與回應行為，與 Eckel et al.（2002）所蒐集的人類實驗數據進行比較，針對「公平分配比例」與「拒絕率」兩項核心指標進行初步對齊分析。





初步結果顯示，大型語言模型（LLMs）在最後通牒遊戲中的行為模式，與人類行為呈現部分對齊。

在提議公平分配的機率上，LLMs 的表現與人類相似，這或許表明其能理解並內化「公平」的概念。可能歸因於其訓練數據中包含了大量人類社會互動的文本，使其習得這種傾向。然而亦可觀察到在特定情境下，LLMs 提議平等分配的機率高於人類，這或許暗示模型更為保守且過度符合社會規範。而關於回應者的行為，由於 LLMs 所模擬的提案多屬公平分配，因此回應者拒絕率相對較低，初步階段尚難評估其真實的拒絕傾向。

此外，Eckel et al.（2002）所揭示之性別與親社會傾向等人口特徵對行為的影響，也將於後續模擬中納入，作更進一步的模擬分析。