Toy Model：可被驗證的「AI 理解人類動機」簡化問題

在我的 final project 中，我提出了 20 年後 AI 可能具備的能力：

能理解人類的動機、情緒與價值，並以主動協作的方式與人互動。然而現在要做到這種「懂人」的智能仍非常困難，因此本報告嘗試設計一個當前可以實作的簡化模型（toy model），作為朝向該能力的第一步。

一、 問題背景：為什麼需要 Toy Model？

AI 目前雖然能生成語言與做決策，但它們仍無法真正理解「人類的動機與心理狀態」——例如焦慮、逃避、遲疑、壓力等因素。

若要讓未來 AI 成為可信賴的合作夥伴，第一步必須讓模型讀懂多模態人類訊號（語音、臉部、文字）並做出具同理心的回應。

因此，我提出的 Toy Model，是要打造一個「可測試、可驗證、可優化」的小型互動任務，讓 AI 在現階段就能開始朝最終能力前進。

二、 Toy Model 任務設計：多模態 × 強化學習 的 AI 對話輔助器

1. 任務概念：簡化版「理解 + 互動」能力的最小原型

我設計的 toy model 是一個 AI 對話輔助任務（conversation-assist task）。

AI 與使用者進行多輪對話，它必須：

- 解析多模態訊號（語音、表情、文字）
- 推測使用者情緒狀態
- 選擇最有可能讓情緒改善的回應策略

2. 方法：使用目前可行的 ML 技術實作

(1) 多模態監督/自監督學習（模仿人類感知）

- Cross-attention、多模態 transformer

- CCA 建立共同情緒子空間
- VAE/ELBO 生成情緒 representation
- InfoNCE 將同一情緒的多模態資料拉近

(2) 強化學習（學會如何回應）

AI 在每輪對話後收到 reward：

- 語氣放鬆 → +1
- 緊張上升 → –1
- 對話中斷 → –2

以及更多相關判斷。

技術：PPO、Actor-Critic、偏好學習、POMDP、OPE/DR

(3) 因果、穩健性與資料治理

- 因果推論避免把相關當因果
- Domain adaptation（MMD/CORAL）
- 差分隱私 DP-SGD、聯邦學習
- 公平性 Equalized Odds / Demographic Parity

三、 評估方式：如何確認 Toy Model 成功？

一、 情緒改善分數

二、 互動滿意度（主觀 + 留存度）

三、 第三方評估（外部觀察者判斷 AI 是否具理解能力）

四、 Toy Model 的代表性與意義

此模型雖然簡化，但它完整呈現：

- 理解人類動機（多模態理解）
- 主動協作（RL 決策）

- 真實可部署要求（因果性、隱私、穩健性）

若 AI 能在此模型中表現良好，就代表它成功邁向「懂人」的智能。