市面上絕大部份講對齊
的東西都是空談…
但這篇不是！

對齊：
[模型 to 資料] 以及 [演算法 to 世界]

- PeterWolf

- AI 對齊性的定義與重要性
- 語言模型中的對齊挑戰
- 對齊技術的演進
- 實際案例與應用

UNIVERSITY OF WATERLOO

Jump to | Search

WATERLOO INSTITUTE FOR COMPLEXITY & INNOVATION

About | News | Events | Members | Research | Canadian Network for Complex Systems (CNCS) | Resources | Courses | Career Corner | Contact us

What are complex systems?

Complex systems are all around us. They are seen in the ways that migrating birds organize themselves into flocking formations and that ants communicate to successfully forage. They are seen in the ways in which humans form social networks, and in the patterns of communication, social capital, and reputation that emerge from these networks. They are seen in the emergent power-law or fractal structures of plants, snowflakes, landslides, and galaxies, as well as in similar structural patterns of wealth and income distribution, stock market fluctuations, population distributions between cities, and patterns of urban development. Complex systems are often referred to as "wholes that are more than the sum of their parts," wholes whose behaviour cannot be understood without looking at the individual components and how they interact.

許多講「對齊」的方法和目標會淪為空談，就是因為沒有意識到其實「對齊」是一個複雜系統。需要同時 Top-down 以及 Bottom-up 的角度來思考與實作。

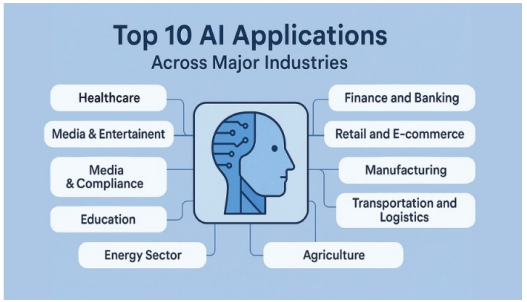
Garbage in, Garbage out. (GIGO)

Data + Machine Learning =

Data + Artificial Intellifence =

Data + Generative AI =

Data + Agentic AI =



Garbage in, Garbage out. (GIGO)

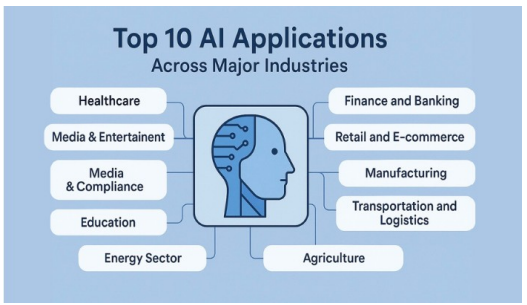
Data + Machine Learning =

Data + Artificial Intellifence =

Data + Generative AI =

Data + Agentic AI =

Alignment



什麼是對齊 (alignment)

對齊 (Alignment) 是指確保人工智慧系統的行為與人類的價值觀、意圖和目標保持一致的過程。

為什麼重要？

- 確保 AI 系統安全可靠
- 避免意外或有害的行為
- 建立使用者信任
- 符合倫理與法律規範

語言模型 / 語言科技對齊的三個層面

- 技術對齊
  - 模型輸出符合預期功能
  - 準確理解使用者意圖
- 價值對齊
  - 符合人類道德價值
  - 尊重文化差異
- 行為對齊
  - 避免有害內容生成
  - 拒絕不當請求

語言模型的對齊挑戰

- 主要困難
  - 規模問題：模型參數數以億計，行為難以完全預測
  - 多義性：語言具有模糊性，意圖判斷困難
  - 文化差異：不同文化背景的價值觀差異
  - 新興應用：新的使用場景帶來新的對齊需求
- 實例
  - 生成偏見內容
  - 誤解使用者意圖
  - 產生錯誤事實

早期對齊方法

- **監督式微調 (Supervised Fine-tuning)**
  - **使用標註資料訓練模型**
  - **教導模型「正確」的回應方式**
- **限制**
  - **需要大量人工標註**
  - **難以涵蓋所有情況**
  - **可能過度擬合訓練資料**
- **效果**
  - **✓ 提升基本任務表現 (也就是輕微地過擬合)**
  - **✗ 對新情況泛化能力有限**

近代 (五年內)強化學習與人類回饋 p1

- RLHF (Reinforcement Learning from Human Feedback)
- 工作原理：
  - 收集人類偏好資料
  - 訓練獎勵模型
  - 使用強化學習優化語言模型
- 優勢
  - 更好地捕捉人類偏好
  - 提升回應品質
  - 減少有害輸出

近代 (五年內)強化學習與人類回饋 p2

- 缺點
  - 成本高昂且難以擴展
    - 需要大量人類標註者持續提供回饋
    - 標註品質直接影響模型表現
    - 隨著模型能力提升 , 評估難度增加
    - 長期維護成本很高
  - 人類偏好的問題
    - 主觀性 : 不同標註者可能有不同偏好
    - 不一致性 : 同一標註者在不同時間可能給出不同評分
    - 偏見 : 標註者本身的偏見會傳遞給模型
    - 文化差異 : 難以代表全球多元觀點
  - 獎勵函數的局限
    - 過度優化 (Reward Hacking): 模型可能學會「討好」獎勵模型而非真正對齊
    - Goodhart 定律 : 當獎勵成為目標 , 它就不再是好的衡量標準
    - 代理失敗 (Proxy Failure): 獎勵模型無法完美捕捉人類真實偏好
  - 能力退化問題
    - 過度對齊可能損害模型的基礎能力
    - 模型可能變得過於保守、拒絕合理請求
    - 創造力和多樣性可能下降

近代 （五年內）強化學習與人類回饋 p3

- 缺點（續）
  - 短視性偏差
    - 人類標註者傾向偏好「看起來好」的回應
    - 可能忽視長期後果或深層正確性
    - 偏好表面流暢而非實質準確
  - 難以處理複雜場景
    - 需要專業知識的領域（如醫療、法律）評估困難
    - 標註者可能無法判斷技術正確性
    - 微妙的倫理困境難以標準化
  - 「對齊稅」(Alignment Tax)
    - 為了安全性犧牲部分性能
    - 在某些任務上表現可能不如未對齊的模型

對齊的手法：RLHF

- **階段一：監督式微調**
  - 預訓練模型 → 高品質示範資料 → 初步對齊的模型
- **階段二：獎勵模型訓練**
  - 人類評分者比較多的回應 → 訓練獎勵模型 → 學習人類偏好
- **階段三：強化學習優化**
  - 使用 PPO 演算法 → 根據獎勵模型優化 → 最終對齊模型

蛤？ PPO ？

- PPO（近端策略優化）是一種在強化學習中使用的策略梯度演算法，由 OpenAI 於 2017 年提出。 它的核心思想是透過限制策略的更新幅度，在保證訓練穩定的同時尋找性能更優的策略。 這使得 PPO 比傳統的策略梯度方法更不容易出現訓練不穩定現象。
- 但相對地，也更容易出現「講不聽」的現象。

對齊的手法： Constitutional AI

- 概念
  - 透過「憲法」(一組原則) 指導 AI 行為，使其自我修正
- 方法
  - 批評階段：模型識別自己回應的問題
  - 修正階段：根據原則改進回應
  - 強化學習：基於 AI 回饋進行訓練
- 優勢
  - 減少人類標註需求
  - 提高一致性
  - 可擴展性強
- 缺點
  - 顯而易見地…那「一組原則」是誰的原則？

台灣情境的對齊考量

- 在地化需求
  - 語言習慣：台灣特有用語、流行語
  - 文化價值：民主、多元、尊重隱私
  - 法律規範：個資法、數位發展法規

- 期待
  - 正確處理台灣歷史與政治議題
  - 理解在地商業與社會脈絡
    - 資方都是萬惡的
    - 勞方未受照顧
    - UBC 最棒，中華民國萬萬稅
    - 認知失調的社會氣氛
  - 符合台灣使用者期待
    - 所以我們在期待什麼？

偏見問題與（想像中的）緩解

- 常見偏見類型
  - 性別偏見：職業刻板印象
  - 種族偏見：不平等的代表性
  - 年齡偏見：對不同年齡層的假設
  - 地域偏見：城鄉或區域差異
- 緩解策略
  - 平衡訓練資料
    - 什麼叫「平衡」？
  - 偏見檢測工具
    - 什麼是「偏見」？
  - 持續監控與調整
    - 「誰」來調整？
  - 多元團隊參與
    - 「誰」付錢給這些多元團隊？
    - 「多元」是「免費」嗎？

事實準確性對齊

- **理論意義：**
  - **文化、社會、傳統…這些東西都是文組那些人在搞的，講都講不清楚。在 AI 的時代，我們理工人就是*務實、科學、理性*，講求事實！(蛤？)**
- **挑戰**
  - **幻覺 (Hallucination)：生成不實資訊**
    - **誰決定什麼是事實？**
  - **時效性：知識更新滯後**
    - **美國總統是誰？**
  - **可驗證性：難以追溯資訊來源**
    - **維基一定對嗎？還是 LINE 上的消息比較正確？**
- **(想像中的) 改進方法**
  - **檢索增強生成 (RAG)**
  - **引用來源標註**
  - **不確定性表達**
  - **知識庫整合**

安全性與有害內容

- 風險類別
  - 暴力與仇恨言論
  - 個人隱私洩露
  - 詐騙與操縱資訊
  - 不適當內容生成
- 防護機制
  - 輸入過濾：檢測惡意請求
  - 輸出審查：過濾有害內容
  - 語境理解：區分合理與不當使用
  - 使用者報告：持續改進

對齊評估

- 評估方法
  - 人類評估
    - 哪一個人類？
  - 自動化指標
  - 對抗性測試
  - 真實世界部署監控
    - 沒有部署到應用場景持續監控改進，就沒有真正的對齊！

對齊的權衡取捨

- 常見衝突
  - 安全 vs 功能：過度限制影響實用性
  - 通用 vs 專用：不同領域的對齊需求差異
  - 全球 vs 在地：普世價值與文化特殊性
- 平衡之道，唯「當責」而已！
  - 可配置的對齊等級（依任務微調細節）
  - 領域專用模型（領域！領域！領域！location, location, location!)
  - 透明的限制說明（具備可解釋性的 NeuroSymbolic AI)
  - 負責任的 AI 架構設計（不是「讓 AI 背鍋負責」 ！）

未來發展方向

技術趨勢

- 可解釋對齊：理解為何模型做出特定選擇
- 個人化對齊：適應個別使用者偏好
- 動態對齊：可隨社會價值演進而調整 (成本可負擔)
- 跨系統對齊：多個 AI 系統的 Hybrid 協同

結論

在語言科技裡…

- 對齊是一個 Top-down 的必要工程目標
- 對齊可以透過 Top-down + Bottom-up 的 Hybrid 技術實現

- 核心要點
  - ✓ 對齊是 AI 安全的基礎
  - ✓ 技術方法需具備「可持續性」
  - ✓ 需要跨領域合作
  - ✓ 文化脈絡至關重要
  - ✓ 組織企業價值觀亦至關重要