1. **您今天上課聽到哪些重點？**

我聽到了很多有關 AI 安全需要考慮的議題, 像是常見的越獄、後門，以及許多沒聽過的名詞，模型飄移、供應鏈攻擊等等，這讓我了解到，AI系統雖然看似穩固，但實際上依舊存在許多被操控或攻擊的可能。

在網路也常看到「如何越獄 ChatGPT」這樣的文章，說一串咒語，好讓使用者繞過 AI預設的道德底線，這樣的新聞之所以存在，或許是因為它觸及到了某種人性，人們幻想著能操作 AI做自己不能做的事情。

在網路行走，本身就像戴了一層面具，更容易放下人們在現實的約束，在此基礎下，一些被社會壓抑、不被接受的念頭，便得到了「像AI傾訴」的這一條道路，可以偷偷的進行。

這樣的新聞存在，並且時不時就會出現，是否代表有一部分人真的幻想能透過 AI 去做現實中不敢或不能做的事? 就像無法親手做出極端行為，便轉而使用AI去做到，這個過程也是一種心理上的出口，對現實不滿的虛擬抒發?而當 AI 走入現實，如果這樣的行為真的存在，是不是會變得無法控制?

**第二題：（必寫）**

1. **請閱讀附件的教材Borg, J. S., Sinnott-Armstrong, W., & Conitzer, V. (2024). “Ch7\_What can we do?” in Moral AI: And How We Get There. Random House，摘要整理裡面的內容（有使用生成式AI協助整理）。**

**為甚麼道德AI難以落實?**

* **技術與倫理脫節：**很多 AI 技術工具雖然能幫助公平性、可解釋性等，但真正怎麼用，需要人做有深度的道德決定，而多數開發者沒有接受過這方面的訓練。
* **角色分工太細、溝通不良：**從設計演算法、建模、產品化到推向市場，每一環節的人背景不同，彼此很少對話，也很少了解整體道德影響。
* **開發文化重效率不重倫理：**「快速迭代」、「先做出來再說」的思維，讓道德問題被推遲處理，甚至被視為阻礙進步的障礙。
* **沒有衡量倫理的工具（指標）：**沒有「衡量是否道德」的標準和數據，公司自然不會投入資源。
* **沒人願意負責倫理問題：**每個人都覺得「這不是我的事」，所以道德責任被稀釋，最終無人負責

**我們可以怎麼做?**

* **擴展與實作道德 AI 技術：**開發更多工具，像是公平演算法、模型說明工具、數據紀錄卡、倫理審查清單……讓它們更好用、更好懂、更容易導入真實場景。
* **改變組織文化與運作方式：**
* 把道德考量納入產品一開始的設計目標
* 鼓勵團隊各角色參與倫理討論，分配道德責任
* 讓高層負起最終責任，基層能放心提出倫理問題
* **提供從業者一輩子的倫理訓練機會：**培養一種叫「系統思維」的能力：了解 AI 是怎麼在社會中運作、會對誰造成影響、如何避免長遠的負面後果。
* **讓全民都能參與 AI 的設計與監督：**

讓一般人也能對 AI 產品發表意見，例如：

* + 開放平台蒐集社會各界回饋
  + 與市民團體合作，了解少數群體的擔憂
* **政府制定靈活的道德 AI 政策：**

讓一般人也能對 AI 產品發表意見，例如：

* + 提供道德AI的標準和獎勵
  + 提供公開、乾淨、可訓練的資料集
  + 支持AI員工的倫理訓練

1. **請說明在閱讀後對於本週第三節課的討論問題有何啟發？**

針對上周同學的回答，我最印象深刻的是對責任歸屬的劃分要怎麼做，同學直接說道「分層分擔」，直言這不應該則怪某個特定的群體，工程師在設計系統應該要思考可能產生的偏見和社會後果，使用工具的人應理解工具的限制，將 COMPAS作為輔助參考…等等。說的激進些，就是「全都有錯」，只不過有分大小，可以說是將可能犯錯的人全部都抓出來數落了一頓，對這樣子的回答，我很認同。

另外有許多同學都有說道 AI 的確會大過去的偏見，因為原理上就是找出符合規律的模型，而所謂規律又是過去帶有偏見所形成的，那麼 AI 自然就會繼承這樣的偏見，這不能怪 AI，只能怪人類。