**1. 工程師如果不能做決定，對理工系所學生的 AI 倫理訓練仍重要嗎?**

重要，就算工程師不能做決定，依舊能透過其他的渠道去發表自己對 AI 倫理的看法，也正因為工程師對 AI 在技術上有更深度的瞭解，發表出的想法通常不會存在對技術上的誤解，對 AI 倫理的了解是有幫助的。

**2. 生成式 AI 所產生的偏誤或偏差，工程師負有責任嗎？**

工程師應該承擔部分責任，因為 AI的偏誤往往來自於訓練數據、模型設計和算法選擇，而這些都是由工程師決定的。但工程師的責任應該有其界限，因為 AI 的行為也受到使用者操作、外部環境等多重因素影響，無法完全由工程師控制。

**3. 需要去調整所訓練的資料嗎？**

是的，應該調整訓練資料，因為 AI 的輸出品質高度依賴於輸入的數據。如果訓練資料存在偏誤，那麼 AI 生成的內容也會帶有相同的偏誤。不同的資料選擇方式，可以降低 AI 偏差，讓 AI更可靠，並減少倫理爭議產生的機率。

1. **您今天上課聽到哪些重點？**

上了一段時間的課程，這堂課總算有比較深入的去了解 AI 是怎麼被建立出來的，有說道 AI 是一個黑盒子，是怎麼被訓練出來的，避免了因為資訊了解不周全導致錯誤的推論結果，老師非常專注於「資料」這件事情上，因為資料的好壞直接影響到了訓練出的 AI 會不會含有人類的偏見。

這就像是在教導一個嬰兒一樣，訓練資料就代表著他的家庭環境，他的身分認同，他的價值觀體系，而他所做出的決策，說出來的話，都建立在訓練資料上，因此，訓練時不應該將良莠不齊的資料通通灌進他的腦袋，而是要透過篩選找出「高品質的資料」，讓 AI 有一顆好腦袋。問題來了，誰來篩選「高品質的資料?」老師就針對這一點講了好多好多東西。

除此之外，就是COMPAS系統的偏誤問題，提到訓練出的機器和人類一樣，有偏見，有提到黑人白人在這個系統的預測下，會因為膚色影響到機器的決策，前面有提到，訓練 AI 就是將資料丟進一個黑盒子，雖說可以選定要怎麼訓練，神經網路要疊幾層，要用甚麼函數影響輸出，但最後訓練出的模型依舊是一團謎，只能通過測試來得到這個模型訓練的好不好，但資料本身背後所隱含的意義包含「黑人的犯罪機率比較高」這件事，自然就會在訓練出來的 AI 就會展現出這樣的偏誤，不能說這 AI 訓練的不好，事實上，它忠實的呈現了從資料當中學到的東西。

**2. 針對各校同學上週第三節所討論的三個問題的回答（請參看Excel檔案or Talk to the City報告），您認為哪些回答您最欣賞或認同？哪些回答您認為需要挑戰？**

* 有許多的回答都延伸到很遙遠的地方，包含
* 避免將 AI 作為唯一的參考方案
* 責任歸誰?
* 不要把機器教壞
* 長者的尊嚴問題 (不感受到被輕視，不感受到被物化)
* 手機的普及化類比到 AI 上面，技術普及化
* 了解 AI 的運作原理，避免誤解
* AI 的環保和維修

這些方向。

我自己的回答相比起來就比較狹隘，畢竟問題問的是「你們是要讓高齡者孤獨到死，還是要製造陪伴機器人陪伴他們?」我的想法首先就肯定是不希望「高齡者孤獨到死」，既然高齡者會孤獨到死，那麼現在這個情況應當是家人大多沒時間或著不在，也因此有了以下十分簡陋的論證，

前提1: 高齡者不應該孤獨到死。

前提2: 陪伴機器人有可能解決孤獨終老的問題。

結論: 綜合前提1, 前提2, 應該製造陪伴機器人陪伴高齡者。

不過，在一提出這個論證之後，就有滿滿的反對聲音自心中浮現

* 陪伴機器人不需要成本嗎?
* 如果照顧的中途出事了怎麼辦?
* 一個冷冰冰的機器照顧自己，真的會解決孤獨的問題嗎?
* 夜深人靜之時，有著完滿家庭、衣食無憂的人都有可能會感受到孤獨感，為甚麼會認為一台機器就能解決這項問題?
* 一個掃地機器人能解決你的孤獨感嗎? (這是來搞笑的選項)

在上周回答問題的時候，我壓根就沒有想那麼多，這週看著其他學校同學們的回答，才意會到了這些沒注意到的方向，這些方向看起來想的太多，太遙遠，僅限理論，會不會成為阻擋進步的絆腳石? 還需要進一步的去思考。

雖說聽了這麼多不同的看法，需要注意的事情有千百項，但「Theory will take you only so far.」我依舊認為這件事情值得一試，就像吃東西一樣，用文字描寫出橘子的甜味，肯定要比實際吃下橘子能感受到的少，真理自會在實踐的道路緩緩浮現 (天啊好中二我好喜歡)。

**3. 如果您剛好有機會校正COMPAS系統在預測上變得比較公平，請以溫育瑋老師提供的一群醫生討論的圖片校正考量為例，以下哪種校正考量較為合理？**

Equal probability

(There are 5 races in the U.S., so 20% of the doctors should be Black.)

這是非常粗淺的方法，跟齊頭式平等差不多，只是基於「種族平均分配」的概念，沒有考慮到現實的族群分布，這個方式會導致一些比例不自然，不一定能解決偏見問題。

Equal representation

(There are 14% of people are black in the U.S., so 14% of the doctors should be Black)

這種方式考慮了美國人口中黑人所占的比例，確保醫生群體的種族分佈與整體人口分佈相符。如果 COMPAS 的預測因為族群差異而不公平，這個方法有辦法幫助降低系統偏誤。

Equal “qualified” representation

(There are 6% of doctors are black in the U.S., so 6% of the doctors should be Black)

這個方法根據現實的醫生比例，確保分配和現狀一致。  
但如果現狀本身就有受到歷史和制度的影響，可能會延續現有的不公平狀況。

No mitigation

(Let the algorithms and data work without human intervention)

這代表讓 COMPAS 繼續使用原始資料預測，不做任何調整，因為歷史上的偏見已經影響資料，可能會導致系統繼續產生對某些群體不公平的預測結果。