主導課程二：人工智慧倫理 (AI Ethics)

課程基本資料

開設學校：東海大學

開授教師： 甘偵蓉

開課級別：學士班

授課語言：中文

課程概述

本課程旨在帶領學生認識及反思AI這項技術及其應用所涉及的倫理、風險與社會議題。首先，將簡介學習機器學習AI的發展歷史以及國際倫理規範，以及哲學倫理學的基本概念。其次，將說明AI的資料來源和分類如何影響AI的預測或決策。接著，將探討AI模型和演算法的設計，並分析AI技術所涉及的社會性與政治性等問題。另外，本課程亦討論AI應用之後所帶來的人類生存危機感、偏誤與歧視、加劇既有的社會不平等、性別刻板印象、以及勞動產業鍊及剝削等倫理與社會爭議。最後，本課程期待修課學生能針對目前已知發生倫理社會爭議的AI專案提出可行的解決方案。

參考書目

1. Borg, J. S., Sinnott-Armstrong, W., & Conitzer, V. (2024). *Moral AI: And How We Get There*. Random House.

2. Russell, S., & Norvig, P. (2021). Chap. 1 & Chap. 27. In Artificial intelligence: A modern approach (4th ed.). University of California, Berkeley.

3. Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds & Machines*, 30, 411–437. https://doi.org/10.1007/s11023-020-09539-2

4. Russell, S. (2019). Chap. 7 & Chap. 10. In *Human compatible: Artificial intelligence and the problem of control*. Penguin.

5. Vallor, Shannon (2016). *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting*. New York, NY: Oxford University Press USA.

6. 凱特.克勞馥(Kate Crawford) (2022)，第二、三、四章，出自《人工智慧最後的秘密》，臉譜文化，pp. 71-108。

7. 泰娜.布策(Taina Bucher) (2021)，第ㄧ、二、四章，出自《被操弄的真實：演算法中隱藏的政治與權力》，台灣商務印書館。

8. 維吉尼亞.尤班克斯(Virginia Eubanks) (2022)，第三、四、五章，出自《懲罰貧窮》，寶鼎出版。

9. 凱西.歐尼爾(Cathy O’Neill)(2109)，第一、十章，出自《大數據的傲慢與偏見》，大寫出版。

10. 約蘭德(Yolande Strengers) &珍妮.甘迺迪(Jenny Kennedy) (2023)，第一、二章，《智慧妻子》，陽明大學交通出版社。

11. 尼克.伯斯特隆姆(Nick Bostrom)(2016)，第十三、十四、十五章，出自《超智慧》，八旗文化。

12. 甘偵蓉(2023)，〈人工智慧科研倫理與風險之基本認識〉，《科技、醫療與社會》季刊，37: 167-220。

13. 甘偵蓉(2024)，〈AI開發過程的倫理權衡：自駕車決策案例研究〉，中研院《歐美研究》季刊第54卷第1期，頁1-68。DOI: https://doi.org/10.7015/JEAS.202403\_54(1).0001 (THCI/TSSCI)

14. 甘偵蓉(2024)，〈人工智慧系統應該內建倫理嗎？人工道德行為者之探討〉，《危機時代的哲學─「後」疫情時期的反思》，五南圖書出版。ISBN: 978-626-393-008-7

課程內容大綱

|  |  |  |  |
| --- | --- | --- | --- |
| **週次** | **日期** | **課程內容** | **備註** |
| **1**  | 114/2/19  | 課程與AI發展史簡介  | 測試Slido, TttC  |
| **2**  | 114/2/26  | 倫理基本概念與AI國際倫理規範簡介  | 測試Slido, TttC  |
| **3**  | 114/3/5  | AI資料的來源與分類之倫理議題  | 講授2節，實體與線上同步討論1節  |
| **4**  | 114/3/12  | AI模型的倫理與社會議題  | 講授2節，實體與線上同步討論1節  |
| **5**  | 114/3/19  | AI演算法的社會與政治性之議題  | 講授2節，實體與線上同步討論1節  |
| **6**  | 114/3/26  | AI演算法的社會與政治性之議題  | 講授2節，實體與線上同步討論1節  |
| **7**  | 114/4/2  | 春節溫書假  |
| **8**  | 114/4/9  | AI與人類生存危機？  | 講授2節，實體與線上同步討論1節  |
| **9**  | 114/4/16  | 期中考週  |
| **10**  | 114/4/23  | AI的部署與運作之倫理與社會議題：偏誤、偏見與歧視  | 講授2節，實體與線上同步討論1節  |
| **11**  | 114/4/30  | AI的部署與運作之倫理與社會議題：自動化不平等  | 講授2節，實體與線上同步討論1節  |
| **12**  | 114/5/7  | AI的部署與運作之倫理與社會議題：自動化不平等  | 講授2節，實體與線上同步討論1節  |
| **13**  | 114/5/14  | AI的部署與運作之性別議題  | 講授2節，實體與線上同步討論1節  |
| **14**  | 114/5/21  | AI的勞工與產業鍊之社會議題  | 講授2節，實體與線上同步討論1節  |
| **15**  | 114/5/7  | 線上展覽分組簡報與同儕互評  | 預錄1分鐘專案自介  |
| **16**  | 114/6/4  | 繳交分組書面報告  | 授課教師與盟校教師評分  |

成績評量方式

|  |  |  |  |
| --- | --- | --- | --- |
| 序號 No.  | 評分項目 Assessment Item  | 配分比例 Percentage  | 相關說明 Description  |
| 1  | 閱讀筆記與上課學習單  | 30 %  | 當天上完課後48小時內上傳  |
| 2  | 課堂參與討論  | 30%  | 依據線上提問或公共討論系統的發言紀錄  |
| 3  | 期中考試  | 15 %  | Take home exam 24小時  |
| 4  | 期末分組書面報告  | 25 %  | 運用上課所學從目前已知發生倫理爭議的AI專案中提出可行解方。 學生互評佔10%，授課與盟校教師評分佔15%  |