「人工智慧倫理」課程計畫大綱

<table>
<tr><td rowspan="2">課程名稱</td><td>中文名稱</td><td colspan="3">人工智慧倫理 I：倫理批判思考工具到公平</td></tr>
<tr><td>英文名稱</td><td colspan="3">AI Ethics I: From Ethical Critical Thinking Tools to Fairness</td></tr>
<tr><td>授課年段</td><td colspan="2">一年級至三年級</td><td>學分數</td><td>2 學分</td></tr>
<tr><td>課程屬性</td><td colspan="4">■專題探究 ■跨領域/科目專題 ■跨領域/科目統整□實作（實驗）<br>□探索體驗 □第二外語 □本土語文□全民國防教育<br>□職涯試探 □通識性課程 □大學預修課程 □特殊需求 □其他_____</td></tr>
<tr><td>師資來源</td><td colspan="4">□校內單科 □校內跨科協同 □跨校協同 ■外聘（大學） □外聘（其他）</td></tr>
<tr><td>學習目標</td><td colspan="4"></td></tr>
<tr><td rowspan="4">課程大綱</td><td>次序</td><td>單元/主題</td><td colspan="2">內容綱要</td></tr>
<tr><td>1</td><td>課程介紹</td><td colspan="2">■ 簡介課程：課程使用工具、操作練習<br>• 教學核心與哲學：以「蓋吉斯之戒」隱喻AI技術的強大與隱蔽性，將倫理視為AI素養的核心，旨在訓練學生的「倫理推論能力」，引導其在具備技術力量時能進行正確的價值抉擇。<br>• 課程架構與內容：由淺入深涵蓋論證辨析基礎、演算法偏誤、資安隱私及生成式AI爭議，延伸至AI伴侶與民主衝擊等特定議題，結合哲學理論分析真實的科技社會問題。<br>• 創新審議與思辨：導入Talk to City (T3C) AI審議工具進行跨校意見整合，透過螺旋式訓練要求學生分析多元觀點並找出反方最強論點，達成深度的人文與科技跨域對話。</td></tr>
<tr><td>2</td><td>基礎篇</td><td colspan="2">■ 機器學習、生成式 AI 基本原理<br>• 釐清技術層級關係：建立人工智慧、機器學習、類神經網路及深度學習間的層次結構感，並認識卷積神經網路（CNN）與大語言模型（LLM）等主流模型類別。<br>• 掌握模型訓練本質：透過線性回歸實例理解資料規律、參數調整及誤差最小化的運作原理，並體會模型複雜度（參數數量）與預測精準度之間的關聯性。<br>• 解構生成式預測機制：理解大語言模型基於「Token」出現機率進行文字接龍的統計本質，強調其結果並非經由嚴謹邏輯推理，藉此釐清AI幻覺產生的根源。</td></tr>
<tr><td>3</td><td>基礎篇</td><td colspan="2">■ 機器學習、生成式 AI 基本原理<br>• 事實與規範的辨析（Is/Ought）：訓練區分「事實描</td></tr>
</table>

| | | | |
|---|---|---|---|
| | | | 述」與「價值評價」，覺察語言中偽裝成事實的主張，避免在討論爭議時從現狀事實直接跳躍至不理性的判斷。<br>• 相關性與因果關係認知：探討統計上的相關性不等於邏輯上的因果關係，分析過度依賴相關性如何導致醫療、法律或社會決策上的倫理偏差與偏誤（Bias）。<br>• 建立主體責任與應對：針對數據抓取、隱私保護與學術誠信等場景，探討在「多手問題」下人作為決策主體的責任歸屬，強調倫理應作為超越法律合規的實踐工具。 |
| | 4 | 基礎篇 | ■ 實踐倫理批判思維工具(一)<br>• 界定規範性論述： 學習識別語句中明示或隱含的評價（如「應該」、「對錯」），區分純粹的事實描述（Is）與帶有價值判斷的規範主張（Ought），並理解倫理並非全然主觀，而是可經由理由支持與邏輯評估的論述。<br>• 分析偽裝的主張：練習察覺語言中的隱性評價，探討當代科技描述如何透過「偽裝成事實的規範主張」影響對AI爭議（如著作權與數據抓取）的判斷。 |
| | 5 | 基礎篇 | ■ 實踐倫理批判思維工具(一)<br>• 釐清規範位階與層次：區分法律（合法與非法）、治理（軟性指標與政策程序）與倫理道德的差異，探討AI倫理如何填補法律落後技術發展時的規範真空，並強調「合規僅是起點，超越合規才是倫理的真正工作」。<br>• 應對當代AI的核心挑戰：針對深度學習的「不透明性（黑盒問題）」、「大規模系統性影響」以及「自體目標缺乏（Autonomy 辨析）」等特性，分析責任歸屬在「多手問題（Many Hands）」下的困難。 |
| | 6 | 基礎篇 | ■ 實踐倫理批判思維工具(一)<br>• 應用十種核心倫理價值：學習在具體場景中應用自由、透明性、公平正義、不作惡、克責性等十種AI倫理原則，將抽象價值轉化為具體的規範要求，例如要求開發者事前公開演算法錯誤率以達成公平。<br>• 建構具體的行動主張：練習精確界定「行動主體（誰）」在特定「情境（何時）」下「應該/不應該（規範詞）」執行「何種行為（做什麼）」，以此對著作權、學術誠信、深偽技術及過度依賴等爭議 |

| | | | |
|---|---|---|---|
| | | | 提出結構化的倫理提案。 |
| | 7 | 基礎篇 | ■ 實踐倫理批判思維工具(二)<br>• 辨析事實與規範：深入區分「事實描述」與「規範主張」，學習辨識語境中隱含價值評價的偽裝主張。<br>• 釐清規範界線：探討 AI 領域中「倫理（道德正當）」、「治理（政策標準）」及「法律（合法底線）」的定義與交互關係。 |
| | 8 | 基礎篇 | ■ 實踐倫理批判思維工具(二)<br>• 後果論與義務論：學習以「後果論」評估AI對整體社會福祉的影響，並以「義務論」檢視基本權利、知情同意與行為正當性。<br>• 德性論與分配正義：解析「德性論」對科技企業與開發者品格之要求，並以「分配正義」檢視AI系統於不同族群間利益與風險分配的公平性。 |
| | 9 | 基礎篇 | ■ 實踐倫理批判思維工具(二)<br>• 建構AI規範論證：實作將「事實證據」結合「倫理價值理由」，推導出利害關係人應採取「具體行動」的完整論證結構。<br>• 在地案例批判與反思：將理論應用於台灣或亞太地區真實AI爭議案例，並訓練使用生成式AI輔助分析時的批判性思維。 |
| | 10 | 課程應用互動、回顧與時事討論 | |
| | 11 | 基礎篇 | ■ 技術的政治意涵<br>• 技術設計的價值內建：透過美國長島公園低矮橋樑的經典案例，解析設計者如何將特定社會階級的偏好與價值觀隱蔽地明刻（inscribed）於物理建築與技術規格中，說明「技術從來不是中立的」。<br>• 理解科技的權力展現：探討科技的政治性體現在三個層面：權力架構的翻轉（如自動化設備將權力移轉給工程師）、基礎設施的強制性權威（如核能與太陽能對極權或分權管理的預設），以及刻意的技術安排如何分配社會的利益與負擔。 |
| | 12 | 基礎篇 | ■ 技術的政治意涵<br>• AI 開發的結構性排除：以Amazon的AI履歷評分系統為例，分析該模型如何因為過度學習科技業「以男性為主」的歷史資料，進而將社會既有的性別不平等「內建」於演算法中，造成系統性地給予女性較低分的偏誤。<br>• 倫理論證的實踐與批判：學習針對AI系統的設計選擇進行「四個規範性追問」：設計目的為何、背後 |

| | | | |
|---|---|---|---|
| | | | 有何規範預設、誰的利益被結構性排除，並運用後果論、義務論或分配正義等倫理觀點，提出反駁與改善的規範論證。 |
| | 13 | 核心主題篇 | ■ AI責任歸屬問題<br>• 事實與規範的辨析： 訓練學生分辨事實證據（P1）與規範理由（P2），並理解如何結合兩者建構出有力的規範性反對意見，避免從「實然」直接跳躍至「應然」的推論謬誤。<br>• AI 系統性偏誤探討： 透過 Amazon 履歷評分系統的案例，分析「不平衡的歷史資料」如何導致演算法對女性求職者產生系統性歧視，並探討科技公司在社會結構失衡下，是否應承擔校正偏誤的道德義務。 |
| | 14 | 核心主題篇 | ■ AI責任歸屬問題<br>• 群體預測與個體權益的衝突：探討英國A-level大學入學成績演算法爭議，分析以「學校過去歷史成績」限制「當屆個別學生排名上限」，如何為了維持程序公平與跨年比較，卻犧牲了弱勢學校學生的實質公平。<br>• 批判性論證與換位思考：進行實務演練，要求學生站在不同利害關係人（如考試監管機構、弱勢學生或教育決策者）的立場，建構或反駁AI系統設計的規範性理由，並釐清在「多手問題」下AI系統決策的實質責任歸屬。 |
| | 15 | 核心主題篇 | ■ AI偏誤問題與公平性<br>• 資料與模型偏誤的生成機制：解析機器學習如何從不平衡的歷史資料（如性別或種族比例懸殊）中學習模式，進而探討偏誤不僅是資料問題，更牽涉到特徵工程、模型選擇及人類決策介入的每一個環節。<br>• 預測效能指標背後的價值取捨：透過生動的案例（如颱風假預測、醫療瑕疵檢測等），說明 Precision（精確率）、Recall（召回率）及 F1 Score 之間存在的權衡關係（Trade-off），並探討在不同情境下，追求單一指標如何隱含特定的倫理與價值選擇。 |
| | 16 | 核心主題篇 | ■ AI偏誤問題與公平性<br>• 預測模型在社會應用上的公平挑戰：以美國司法再犯風險評估系統（COMPAS）為例，探討當演算法預測結果對特定弱勢群體（如黑人）產生系統性高估風險時，所引發的分配正義、程序正義及間接歧 |

| | | | |
|---|---|---|---|
| | | | 視等倫理爭議。<br>• 修補偏誤與反映現實的兩難：反思當技術工程師面對帶有偏見的現實資料時，應採取何種「人類介入（Human in the loop）」策略；探討AI系統應「忠實反映不完美的現實」，還是該被設計為「朝向公平與理想目標校準」的規範性問題。 |
| | 17 | 分組報告 | ■ 期末分組口頭報告、期末倫理審議報告 |
| | 18 | 分組報告 | ■ 期末分組口頭報告、期末倫理審議報告 |
| 學習評量 | 一、課堂參與討論（含出席）(50%)<br>• 每週設計題目並結合小組討論時間，鼓勵學生交換觀點與意見，打破原有倫理邏輯。<br>• 根據討論內容與完成度進行評分，鼓勵學生主動開口發言。<br>二、書面作業（20%）<br>• 課程基礎篇、核心主題篇，核心設計2次書面作業。<br>• 透過作業完成度、內容精確程度進行評分。<br>三、期末（小組）報告（30%）<br>• 以個人方式進行期末專題，主題須包含AI倫理相關議題。<br>• 評量包含問題定義、技術應用、創意表現、資料分析、成果呈現與簡報表達。<br>四、教師自主與彈性評量原則<br>• 本課程提供建議成績表評定標準，提供高中端授課老師參考。然最終成績評定仍以授課教師之專業判斷為準，教師可依學生實際表現、學習投入度與課堂互動等綜合因素進行調整。<br>• 鼓勵各班老師依教學現場實況，自主評量與賦分，以確保評量結果貼近學生真實學習狀況。 | | |
| 備註 | | | |