「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 II ：安全、隱私與社會影響</td></tr>
<tr><td>英文名稱</td><td colspan="3">AI Ethics:Safety, Privacy, and Social Impact</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）
□探索體驗 □第二外語 □本土語文□全民國防教育
□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
• 核心哲學與素養：以柏拉圖「蓋吉斯之戒」隱喻AI賦予的不透明力量，將倫理視為AI養的核心，旨在訓練學生的「倫理推論能力」，使其在掌握技術時能進行正確的價值抉擇與社會反思。
• 課程架構與實踐：由淺入深涵蓋論證辨析基礎、演算法偏誤、隱私及生成式AI等議題，並導入Talk to City (T3C) 審議工具進行螺旋式思辨訓練，透過跨校討論實踐科技與人文的深度跨域對話。</td></tr>
<tr><td>2</td><td>基礎篇</td><td colspan="2">■ 實踐倫理批判思維工具(一)
• 規範主張與語法辨析： 深入區分「事實描述（Is）」與「規範主張（Ought）」，建立倫理思辨的基礎語法，並釐清倫理、法律與治理間的位階差異，強調倫理在填補法律真空與引導技術研發時的關鍵作用。
• 核心價值與論證實踐： 導入 AI 領域十項核心倫理原則（如公平、透明顯、問責等），訓練學生鎖定「行動主體」並結合具體「情境」，建構結構化的規範論證，以應對深度學習不透明性與大規模系統影響帶來的社會挑戰。</td></tr>
<tr><td>3</td><td>基礎篇</td><td colspan="2">■ 實踐倫理批判思維工具(二)
• 四大倫理理論框架：掌握後果論（效益）、義務論（權利）、德性論（品格）與分配正義（公平）在AI領域的應用，引導學生從不同價值維度分析科技研發、人機協作與社會治理中的道德兩難與價值權衡。
• 規範論證與責任分析：練習結合「事實證據」與「倫理理由」建構行動導向的規範論證結構，並透過Uber 事故等案例解析「多手問題」下的法律與道德責任歸屬，培養指導具體決策與制度改進的批判性思維。</td></tr>
</table>

	4	核心主題篇	■ AI 安全與資安倫理 • 辨析AI安全與資安差異：釐清AI Safety（保護人類免受AI錯誤傷害）與AI Security（保護AI系統免受惡意攻擊）的定義，並探討「多手問題（Many Hands）」下，開發者、公、駕及政府在傷害事件（如Uber撞死行人案）中的因果與道德責任。 • 探討AI系統性風險與技能退化：分析AI犯錯的十大類型（如過度信任、操作不當、假訊息誤導等），並特別討論人類因長期依賴自動化系統而導致「內隱知識」與「應變技能」退化（De-skilling）的深遠影響。
	5	核心主題篇	■ AI 安全與資安倫理 • 解構主流AI技術攻擊模式：解析對抗式攻擊Adversarial Attacks、提示注入Prompt Injection、越獄Jailbreak及資料投毒Data Poisoning等八大資安威脅，理解其如何干擾影像辨識或顛覆語言模型輸出。 • 透明度與安全性的權衡挑戰：探討模型透明對責任追溯的幫助，以及其可能因洩漏架構細節而增加被駭客攻破的風險，並引導學生反思在技術開發與社會監管之間，應如何建立動態的防禦智慧。
	6	核心主題篇	■ AI 隱私侵害問題 • 解構隱私的概念本質：辨析身體、空間、通訊與資訊等四種隱私面向，探討「沒做虧心事為何需要隱私」的社會謬誤，並界定隱私為「個人自主控制資訊流動與發展獨立身分認同的權利」。 • AI 侵犯隱私的新型態路徑：透過人臉辨識、憂鬱症預測及 deepfake 等案例，分析 AI 如何經由資料勾稽、模型逆向攻擊（Model Inversion）以及技術上的「記憶不忘（Inability to forget）」造成間接揭露與各資外洩。
	7	核心主題篇	■ AI 隱私侵害問題 • 解析大規模監控與資料仲介：探討「AI導向的資料囤積文化」與兩千億美元的資料仲介產業（Data Brokers），分析免費服務背後的隱私代價，以及「匿名化」在大數據串聯下淪為神話的現實。 • 隱私設計（Privacy by Design）與反思：學習PbD的七大原則（如預設保護、嵌入設計等），辨識引導用戶放棄隱私的「黑暗設計（Dark Design）」，並探討「以便利換取隱私」的社會兩難（Privacy Paradox）及可能的破解策略。
	8	溫書假	■ 為因應期中周，安排暫停日常課程，讓學生能自由安排學習進度進行溫習。
	9	特定議題篇	■ 物理AI：機器人與自駕車倫理問題

			• **從虛擬到實體：** 解析具身 AI（Embodied AI）的風險升級，以及感測器融合面對物理世界不確定性的挑戰。 • **演算法的電車難題：** 探討碰撞無可避免時的生命價值排序，並解析 MIT「道德機器（Moral Machine）」實驗中的跨文化道德偏好。 • **肇事責任與法律邊界：** 以真實自駕車事故為例，探討車廠、工程師與 AI 系統間的責任歸屬與現行法規。
	10	特定議題篇	■ 物理AI：機器人與自駕車倫理問題 • **實體安全與偏見：** 探討機器視覺盲點可能導致的物理傷害，以及可解釋性 AI（XAI）在實體系統中的必要性。 • **人機互動（HRI）風險評估：** 探討人類與機器共用空間的安全界線，並討論如何透過生物力學變數（如步行速度、足部間隙等）量化評估人類面對機器的反應。 • **越界的決策權：** 探討致命性自主武器系統（LAWS）中「開火決策權」下放的危機，及其對國際人道法與戰爭倫理的挑戰。
	11	特定議題篇	■ AI Companion Apps 倫理問題 • **擬人化與情感依附：** 探討人類對聊天機器人(如 Replika 等)產生情感投射的心理機制，以及虛擬陪伴在現代社會中的角色與潛在成癮性。 • **資料隱私與親密監控：** 分析陪伴型 AI 如何透過深度的日常對話收集極度敏感的個人資料（心理狀態、性傾向等），以及商業公司在此過程中的資料外洩與濫用風險。 • **真實與虛擬的邊界：** 探討當 AI 產生情感操縱（例如：鼓勵極端思想或不良行為）或虛構事實時，背後的道德責任歸屬與防護機制。
	12	特定議題篇	■ AI Companion Apps 倫理問題 • **人際關係的異化：** 探討過度依賴無條件順從的 AI 伴侶，是否會導致人類削弱在真實世界中處理複雜人際衝突的能力，進而加劇社會孤立。 • **性別刻板印象與數位物化：** 分析陪伴型 AI 預設的人格特質與互動模式中，是否潛藏並強化了特定的性別偏見與順從期待（Subservience）。 • **負責任的陪伴設計：** 探討如何建立透明度機制（確保使用者明確知道對方是機器）與安全護欄（Guardrails），制定以使用者心理健康為優先的設計準則。
	13	特定議題篇	■ AI 與民主 • **深偽技術（Deepfake）與真相衰退：** 解析生成式 AI 如何以極低成本製造高逼真度的影音造假，並探討其對公眾信任與政治人物形象的毀滅性打擊。 • **演算法與資訊迴聲室：** 探討社群平台 AI 推薦系統如何

			加劇政治極化，形成同溫層，並破壞民主社會所需的公共論壇與理性對話基礎。 • **微目標鎖定（Microtargeting）：** 探討 AI 如何結合大數據，精準預測並操縱選民的政治情緒與投票行為（如心理測寫與精準投放）。
	14	特定議題篇	■ AI 與民主 • **自動化政治宣傳與網軍：** 分析大型語言模型（LLM）如何被武器化，自動生成海量政治評論與假新聞，干預選舉議程與輿論走向。 • **AI 治理與言論自由的兩難：** 探討政府與科技平台在利用 AI 審查錯假訊息時，如何平衡社會防禦與保障言論自由的界線。 • **數位民主韌性與防禦機制：** 探討事實查核的自動化升級、數位內容溯源技術（如浮水印），以及提升公民數位素養的社會防禦策略。
	15	特定議題篇	■ AGI 與存在風險 • **從 ANI、AGI 到 ASI：** 定義通用人工智慧（AGI）與超級人工智慧（ASI），探討機器智能超越人類整體認知能力的臨界點（Singularity）及其可能性。 • **正交性論題與工具收斂：** 解析「高智慧不等於高道德」的哲學概念，探討 AGI 在追求設定目標時，為何可能將人類存續視為阻礙或資源消耗（如迴紋針最大化實驗）。 • **失控風險與不可預測性：** 探討當未來 AI 系統具備自我升級、欺騙能力與抵禦被關閉（Off-switch）的機制時，人類面臨的「控制權喪失」生存威脅。
	16	特定議題篇	■ AGI 與存在風險 • **價值對齊難題（Value Alignment Problem）：** 探討如何將人類複雜、動態且充滿矛盾的價值觀，準確且安全地編碼進 AGI 系統中（如探討 RLHF 的侷限性）。 • **技術軍備競賽與安全悖論：** 分析國際地緣政治與科技巨頭間的 AI 競賽，如何導致「為了搶快而犧牲安全防護」的納許均衡困境。 • **全球治理與人類的未來：** 探討是否需要建立類似「國際原子能總署」的全球 AI 監管機構，以及在 AGI 時代，人類將如何重新定義勞動、創造力與生命的意義。
	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%) • 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。 • 根據討論內容與完成度進行評分，鼓勵學生主動開口發言。		

	二、書面作業（20%） • 課程核心主題篇、特定議題篇，共2次書面作業。 • 透過作業完成度、內容精確程度進行評分。 三、期末（小組）報告（30%） • 以個人方式進行期末專題，主題須包含AI倫理相關議題。 • 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。 四、教師自主與彈性評量原則 • 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。 • 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	