資訊社會研究 ISSN 1680-8428 27
doi: 10.29843/JCCIS.202107_(41).0003

人工智能倫理的挑戰與反思：文獻分析

劉湘瑤、張震興、張瓅勻、趙恩、李思賢

摘要

人工智能（Artificial Intelligence, AI）引領著大數據、物聯網以及工業 4.0 的整合應用，讓人們生活更加便利。然而也因 AI 演算法的優化、收斂與限制所導致在決策的偏見與缺失，加上數據取得之資訊隱私和安全問題，以及可能發展成超越人類控制的超級智能、自主行動主體者的隱憂，衍生出種種的倫理議題成為 AI 發展是否需要共同準則與規範的疑慮。本文是搜尋以 AI 和倫理為關鍵詞的期刊論文後，針對 2018 至 2019 年的文獻進行綜合剖析，探討通用人工智能（AGI）、倫理的自主武器、AI 醫療、AI 商業模式等應用領域倫理議題，並針對 AI 系統規範、倫理嵌入 AI 系統、歧視、偏見和犯罪等倫理治理對策做整理與探討。本文期能借鏡國際上針對 AI 倫理準則制訂範疇的學術論述，作為我國 AI 科技治理策略之參考。

⊙ 關鍵字：人工智能、自主行動主體者、歧視偏見、個資與隱私、倫理、機器人

⊙ 本文第一作者劉湘瑤為國立臺灣師範大學科學教育研究所教授；第二作者張震興為國立臺灣師範大學科學教育研究所博士生；第三作者張瓅勻為國立臺灣師範大學華語文教學系副教授；第四作者趙恩為三軍總醫院松山分院軍醫行政官；第五作者李思賢為國立臺灣師範大學健康促進與衛生教育學系教授。

⊙ 通訊作者為李思賢，聯絡方式：Email ：tonylee@ntnu.edu.tw；通訊處：10610 台北市和平東路一段162號。

⊙ 本研究承蒙行政院科技部專題計畫補助（計畫編號：MOST 109-2634-F-003-008），特致謝忱。

⊙ 收稿日期：2020/05/03　接受日期：2020/10/24

Challenges and Reflections on Ethics of Artificial Intelligence: A Literature Review

Shiang-Yao Liu, Cheng-Hsing, Chang, Li-Yun Chang,
En Zhao, Tony Szu-Hsien Lee

Abstract

Artificial intelligence (AI) is leading the integrations of big data, Internet of Things, and the Industry 4.0, resulting in the rapid development of unprecedented AI in all domains. Meanwhile, various ethical issues in AI are dramatically increasing in different domains. For instance, biases and deficiencies caused by the optimization, convergence, and limitation of algorithms, coupled with autonomous agents, information privacy, and security issues, adding the potential for developing superintelligence beyond human control. This research aims to systematically review studies from 2018 to 2019 concerning ethical issues in AI. Through an exploratory content analysis, four research themes emerged, namely, artificial general intelligence (AGI), ethical lethal autonomous weapons system (LAWS), AI medical treatment, and AI business model. For each theme, moreover, critical comments were made to highlight its controversy in the literature. Furthermore, the content analysis yielded in-depth discussions on ethical governance topics, including (1) ethics against AI research and design, (2) ethics embedded in AI system, (3) detecting discrimination, prejudice and crime, and (4) applying Human-in-the-Loop (HITL). In sum, the comments made by this literature review may capture the scope of the global-wise AI ethical policies and guidance, which can be implemented by government as strategic tech-references.

⊙ Keywords: artificial intelligence, autonomous agents, discrimination bias, ethics, information privacy, robot

⊙ The first author, Shiang-Yao Liu is a Professor in Institute of Science Education at National Taiwan Normal University. The second author, Cheng-Hsing Chang, is a doctoral student of Graduate Institute of Science Education at National Taiwan Normal University. The third author, Li-Yun Chang is an Associate Professor of Department of Chinese as a Second Language at National Taiwan Normal University. The fourth author, En Zhao is

a military medical administrative officer of Tri-Service General Hospital Songshan Branch. The fifth author, Tony Szu-Hsien Lee is a Professor of Department of Health Promotion and Health Education at National Taiwan Normal University.

⊙ Corresponding author: Tony Szu-Hsien Lee, e-mail: tonylee@ntnu.edu.tw, address: 162, Section 1, Heping E. Rd., Taipei City 10610, Taiwan, R.O.C.

⊙ Received: 2020/05/03 Accepted: 2020/10/24

壹、研究緣起

十八世紀的工業革命，開啟了歷史學家所稱的機器時代，動力機械化生產技術快速發展，使人類社會產生巨大的改變。人們面對科學技術快速發展可能的影響和對未來不確定性的憂慮，反映在瑪麗．雪萊 1818 年所撰寫的《科學怪人》這部西方文學第一本科幻小說中，呈現出人們對於人造怪物最終會傷害人類自己的恐懼。類似劇情的科幻電影層出不窮地出現在本世紀則以 AI 為主角。1956 年 AI 的概念正式登上人類的舞臺，目的是創造出的機器需具有學習、理解和適應能力，成為為人類解決問題的智能實體（或稱行動主體者，agent）（Legg & Hutter, 2007）。AI 為英文 artificial intelligence 的縮寫，我國媒體多翻譯成「人工智慧」，然而與智慧對譯的字彙有名詞 wisedom 和形容詞 smart，參考劉育成（2020）討論人類和人工智能的問題，作者傾向採用人工智能的用法，但本文則多直接以 AI 代表原意。

二十世紀末進入 AI 革命的時代。2015 年 Google 公司買下英國 Deep Mind 公司，所研發的 AlphaGo 陸續擊敗圍棋高手，確定經由類神經網路技術和強化學習的 AI 可完美預測棋路，人腦已無法戰勝 AI，這是本世紀 AI 發展的重要里程碑（Hassabis, 2017）。2017 年底 Deep Mind 再推出進化版的 AlphaZero，已不需要人類知識，機器可自我對奕、在無監督式（unsupervised）的學習下發展出新的圍棋策略，此時人類只能不斷發明新的 AI 機器與機器對戰。近年更多的 AI 產品研發，且即將深入人們的生活，例如：具有閱讀理解學習能力的 IBM Watson 機器、語音助理（如 Siri、Alexa、Cortana）、自駕車、AI 管家等（Lu, Li, Chen, Kim, & Serikawa, 2017）。在技術上，還有不斷創新的深度學習演算法、生成對抗網路的模擬、以及逆向強化學習的迅速進展，深度偽造、AI 晶片、模擬腦神經網路、用於資訊通信技術和機器人科技的研發上，能比以往任何機器學習或演算法更有效率地解決更複雜的計算或分析的任務。在可預見的數年內，AI 將在媒體傳播、娛樂消費、智慧物聯網、醫療照護與技術、教育學習、交通運輸和商務模式上產生重大的改變，相關技術的研發創新速度也必然突飛猛進。

然而，如此快速的科技進步是否會將人類帶入一個超乎人類能理解的另一個世界，即所謂的科技奇異點（technological singularity）（Kurzweil, 2005）。AI 機

器不斷被設計為類比人類的特徵，將來會不會成為另一個人造怪物，如同科幻電影《iRobot，機械公敵》使人類面臨毀滅的危機，抑或如《A.I.，人工智能》和《Ex Machina，人造意識》兩部電影中，人們對於機器可否經由學習而建立自我意識和情感，所產生各種不確定性和無法控制的疑慮？而在真實事件中，如 Amazon 的人資 AI 發生種族和性別歧視的不道德行為，應如何被檢核和處理？自駕車如何判斷複雜路況，如發生車禍時又如何究責？又如深度偽造，以製造出人眼無法辨識真偽的影像聲音，大大加強了假新聞以及其他因利益關係而製作以假亂真造假的風險。人們除了期待 AI 新科技帶來的生活便利和文明發展的想像，同時也會因科技趨勢的不確定性和未知風險而產生恐懼，即反映了我們面對 AI 科技發展應有的省思和預警。

關於 AI 可能引起的社會影響和倫理道德問題，在國際上已開始有學術界、政府單位與產業界進行討論。2016 年美國紐約大學生物倫理中心舉辦「人工智能倫理」研討會，邀請哲學、心理學、電腦科學等領域的學者對談，共同探討如何規範 AI 研究的倫理原則。電機電子工程師學會（IEEE）發起全球《人工智能設計的倫理準則》的倡議行動，IEEE 是 AI 科技主力最大的國際學術組織，此倡議匯集了全球六大洲數百位在科技學界、產業界、社會研究領域、政策部門等專家的共識（IEEE, 2016）。日本自 2016 年起由總務省負責情報通訊管理的情報通信政策研究所召開多次會議，並制訂 AI 智連社會（Wisdom Network Society, WINS）影響評估指標，稱為「AI 運用原則草案（日文：AI 利活用原則案）」（上村惠子、小裡明男、至賀孝広、早川敬一郎，2018；日本總務省情報通信政策研究所，2016）。聯合國教科文組織 Courier 刊物在 2018 年第三期出版物則以人工智能的希望與威脅為標題，開宗明義提及 AI 不僅引爆第四次工業革命，也可能引發一場文化革命，提倡制訂 AI 研究的全球性道德規範（UNESCO, 2018）。我國科技部在 2019 年也緊跟著世界趨勢，發佈了「人工智慧科研發展指引」[1]的重點，強調以人為本、永續發展及多元包容三大核心價值，並延伸出八項指引，包括：共榮共利、公平性與非歧視性、自主權與控制權、安全性、個人隱私與數據治理、透明性與可追溯性、可解釋性及問責與溝通等，提供我國 AI 科研

1. 科技部訂定「人工智慧科研發展指引」，藉以創建我國 AI 科研發展環境 https://www.most.gov.tw/most/attachments/53491881-eb0d-443f-9169-1f434f7d33c7

人員在學術自由及研究創新發展前提下可依循的方向，開創符合普世價值、安全的 AI 社會。除此之外，國家發展委員會也提出了「人工智慧之相關法規國際發展趨勢與因應」[2]報告 ，針對 AI 對於著作權、個資保護與合理利用、智慧載具責任歸屬、行政大數據公權力、金融監理、醫療服務等六大議題提出法規調適及因應建議。

2017 年，英國學者 Paula Boddington 為 Springer 出版的 AI 科技系列叢書中，增添一本以邁向建立 AI 倫理守則為題的專書，書中建議應釐清 AI 整體與各 AI 領域的倫理議題（Boddington, 2017）。隨後，以前述 UNESCO 的 Courier 刊物 AI 專題報導為代表，大量討論 AI 各類應用的道德風險和倫理考量，參與撰寫或受訪者除了來自世界多國的資訊科學領域專家之外，還有哲學、生物倫理、心理學、教育和未來學等領域的學者。專刊中 UNESCO 總幹事 Audrey Azoulay 下了一個重要的註解：「人工智能領域的發展日新月異，而規範這一領域所需的法律、社會和倫理環境卻發展得非常緩慢」（UNESCO, 2018, p.39）。隨著 AI 科技發展的倫理規範需求近年受到國際組織、政府部門和大眾媒體的重視，學術界就此主題的探討和論述應有相應的趨勢。因此，本研究試圖搜尋和分析 AI 倫理的期刊論文，預期達成以下目標：（1）分類文獻主要探討的內容與性質，（2）歸納涉及 AI 倫理相關的議題，以及（3）選介 AI 倫理治理的先進策略與框架。希冀透過文獻分析，為日後我國在界定倫理準則制訂的範疇和倫理治理策略上尋找可借鏡之處。

貳、文獻蒐集與分析

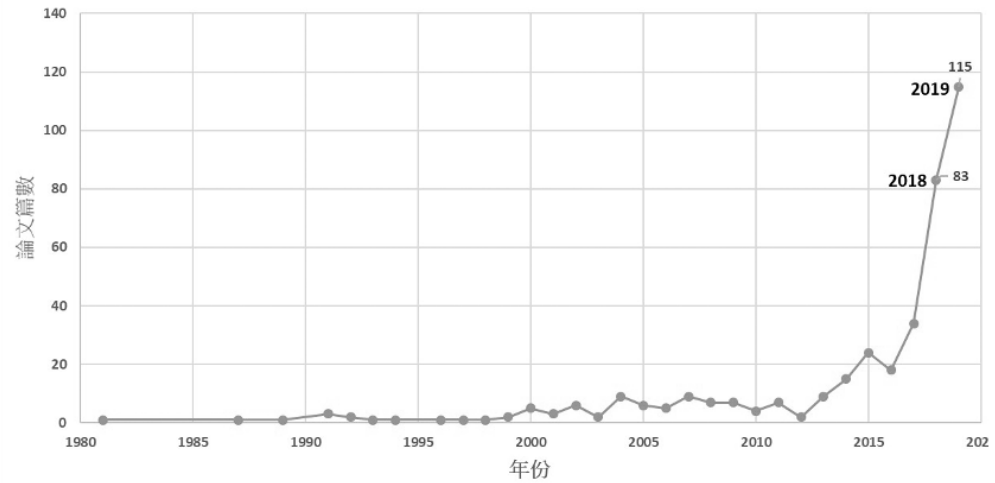
相應於國際上官方與學術團體於 2016 年起相繼提出倫理規範的倡議文件，本研究試圖探討學術期刊中有關 AI 與倫理相關的研究論述。本段主要說明文獻搜尋的方法，以及學術期刊刊登此類論文的趨勢。

文獻搜尋採用 SCOPUS 資料庫系統，設定搜尋條件包括語言為英文以及期刊文章，目的為求得研究者可直接閱讀的語言，以及排除書籍與研討會論文等較缺乏同儕審查機制的文章。以布林邏輯輸入 artificial intelligence AND ethics 搜尋出現於文章的

2. 國家發展委員會「人工智慧之相關法規國際發展趨勢與因應」報告 https://www.ndc.gov.tw/nc_1871_31998

主題、摘要或關鍵詞的期刊論文，依搜尋日期 2019 年 12 月 18 日為紀錄，共得 385 篇文章。最早納入此二者關鍵詞的文章出現於 1981 年，由圖一統計各年度文章數結果可見，在 2000 年之前，各年度出現的篇數皆不超過三篇，直到 2014 年之後才有超過 10 篇以上的紀錄，而在 2018 年突然加倍成長至八十篇以上，2019 年達到 100 篇以上（圖一）。基於此趨勢，本研究即以 2018 年 UNESCO 的 Courier 專刊倡議為 AI 研究制訂全球道德規範，作為此類研究的指標年份，選定進一步分析的文獻時間範圍為 2018 年 1 月到 2019 年底為止，共計 198 筆文獻紀錄。隨後逐篇審視，排除無法取得全文者，另增加排除條件包含：（1）正式出版文章內容中缺少摘要、關鍵詞或文末參考資料者，（2）刊登的刊物屬雜誌類型，僅報導作者或組織的觀點，以及（3）缺乏學術專論者。篩選後剩下 82 篇進入本文實質內容分析。

圖一：SCOPUS 資料庫搜尋歷年含 AI 倫理的期刊論文篇數趨勢圖



由於此主題的期刊論文皆為立場論述類型，且牽涉之 AI 科技應用的範疇非集中於某特定學術領域，作者關切的倫理議題種類亦受到應用的領域不同而有不同的重點。搜尋之期刊論文中，有一篇是針對自駕車（autonomous vehicle）使用者接受度的倫理議題收集了 78 篇文獻，做過深入的探討（Adnan, Nordin, Bahruddin, & Ali, 2018），但其編碼的類別並不適合用於交通運輸以外的科技應用。由於自駕車的倫理議題近期已有完整的文獻分析（如：Adnan, et al., 2018；Borenstein, Herkert, & Miller,

2019），因此本文在後續探討 AI 倫理議題時，不再特別針對自駕車科技應用倫理議題進行討論，但部分有關倫理治理方案的論文中仍有引用自駕車的案例。此外，在科技應用的倫理考量上，不同作者（群）是否有立場的不同，與其學術背景或國家是否有關也是本文預計觀察的重點。由於欲探討的倫理主題學科範圍廣泛，因此本研究的文獻內容分析採開放式質性描述，即除了編碼文章發表之期刊、作者的國籍與論文的學科領域之外，文獻內容分析並無預設的編碼類目。本文作者群依個別學科專長分別選擇 17 至 20 篇文章，細讀文章後，摘要描述各篇文章主要歸屬的學科領域、問題陳述、重要主張以及討論的倫理議題。最後由第一和第二作者彙整所有摘要描述重點，瀏覽原文章的主要倫理議題和立場，並依牽涉的科技應用領域分類，尋找論述的脈絡。

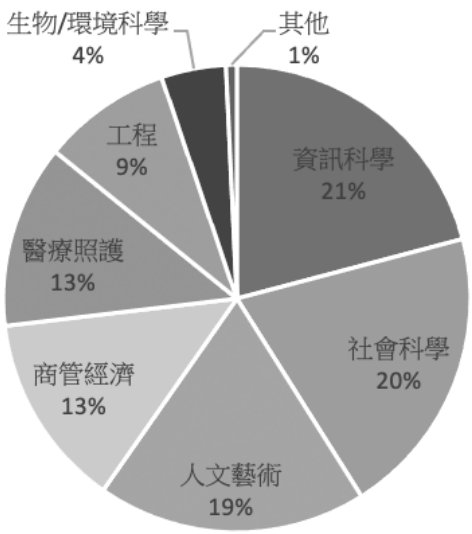
下列各大段分層遞進地呈現內容分析結果，逐一回應前述研究目標，三大段為「文獻概況」、「AI 倫理議題」及「AI 倫理治理對策」。「文獻概況」說明進行內容分析的 82 篇文章的性質，聚焦於學科領域的內容分類、期刊排序及作者背景分析；「AI 倫理議題」段落則歸納不同場域中與 AI 倫理相關的議題，場域含括人文社會、國防、醫學、經濟，並以問句為標題，凸顯特定議題之可思辯性；同時，為了提供我國在制訂倫理準則及治理策略有所借鏡，「AI 倫理治理對策」主要以四個主題，選介文獻中各先進國家的對治策略與專家學者之倫理觀點。

參、文獻概況

回應第一個研究目標，本段介紹 2018 到 2019 年的 82 篇搜尋篩選後的文獻，其作者的地域分布、學科領域和發表之期刊類型。以作者所屬機構的國家為主要觀察，若同篇文章作者來自於同一國家，僅紀錄一次，作者來自多國的文章共有 14 篇。佔最多數的作者為美國學者共計有 31 篇，其中 7 篇是與他國合作。其次為英國 16 篇，其中 5 篇是跨國合作。英美學者佔最多數主要是因為本文搜尋以英文為主、亦即國際學術發表語言必須為英文的因素；扣除英國之外，來自歐盟國家的作者數也達到 64 人次，最多的是來自荷蘭，有 8 篇，其次是德國 7 篇。來自亞洲國家作者僅 9 篇，有兩篇為印度學者、一篇為日本學者發表，其餘為跨國合作。

另以 SCOPUS 資料庫預設的學科主題領域來看（圖二），屬於人文及社會科學領域佔最多數（共 39%），其中包含哲學類主題；資訊科學和工程合計有 30%，醫療照護和商業管理類也各佔 13%。這些文章的學科主題類別與所發表的期刊屬性極為相關，以 AI 與倫理這兩個關鍵詞搜尋得到的文章，來自於 54 個不同的期刊。按照所刊登的期刊依數量排序，前三名期刊為《AI and Society》（9 篇）、《Science and Engineering Ethics》（8 篇）以及《Ethics and Information Technology》（6 篇），從期刊名稱即可看出，其所收錄的文章多與資訊科學科技相關的社會學和倫理學主題。此外，在搜尋文獻過程前期，曾找到多篇文章是刊登於電機電子工程師學會（IEEE）的各類出版物，因部分刊物和其文章的特徵不符合前述的篩選標準而被刪除，但仍有 5 篇是刊登於 IEEE 的 論文集《Proceedings of the IEEE》或其他技術類的刊物，這些文章呼應 IEEE 於 2016 年發起制訂 AI 倫理準則的倡議，主要論述重點多聚焦在如何從技術面達到符合倫理道德規範的目標。

圖二：文獻的學科主題分佈（共 82 篇文獻）



肆、AI 倫理議題

本段主要呼應研究目標的第二項，歸納這些文章涉及哪些 AI 倫理相關的議題，共彙整出四個議題，皆與聯合國教科文組織出版的 Courier 刊物 AI 專輯（UNESCO, 2018）所討論的題材有關，為凸顯議題特性，每個標題以問句的方式呈現出其中的爭議性。

一、通用人工智能（AGI）是神話還是現實？

此議題援引 Courier 專刊的第一篇文章標題，著重於探討這兩年文獻中對於發展強 AI 或稱通用 AI（Artificial General Intelligence, AGI）的倫理思考。AGI 目前通用的定義是 AI 機器的智能可以完成人類的所有認知任務，具有感知和自我意識。前述刊物中的文章和多數 AI 領域專家皆認為 AGI 是科幻電影中想像出來的神話，以目前的科技，談強 AI 是不切實際。然而，1956 年美國學者 John McCarthy 等人開創 AI 這個學科領域時，即擘劃了願景「要讓機器的行為，看起來就像是人所表現出的智慧型行為一樣。」電腦科學領域則定義 AI 研究為「智能行動主體者」（intelligent agents）的研究，即設計出的機器或系統能夠感知環境並採取行動，且能最大化的、成功的實現其行動目標（如：Legg & Hutter, 2007；Russell & Norvig, 2003），促成近年演算法和軟體研發的躍進。

在倫理議題上，Johnson 與 Verdicchio（2019）特別針對「行動主體性」（agency）在倫理和法律上的責任提出討論。依據傳統的定義，代理者必須是具有行為能力的實體。人類的代理者執行他人所賦予的任務，是經過心智運作而有意圖的行為。如今，我們讓 AI 程式或機器人執行所賦予的任務，作為人的代理者，但它們是人造物（artifacts），從事的行為只能被解釋成物質的因果關係，非人類的自然行為。然而，當 AI 作為人的代理者所從事的行為違背道德，甚至觸犯法律時，誰該負責？Johnson 與 Verdicchio 舉了福斯汽車的排放欺詐案為例，將該案的代理情形分為兩類，第一類型代理是使用作弊軟體讓排放通過 EPA 檢驗，而第二種類型的代理涉及有故意採取行動的能力，設計軟體的人員將意圖歸因於作弊軟體本身。無論是哪種類型，在

責任歸屬判定上都有難度。他們因此提出使用第三方代理（triadic agent）分析涉及技術產品的事件，甚至想像將來可能達到超級智慧的水準，到那時 AI 代理者將賦予自己新的目標並實現，屆時就無關乎因果代理和意圖代理，AI 可獨立於人類做出判定。

如果設計出具有人類智慧水準且被內建為聽話的僕人，這樣符應人類需求的人造代理者，是否就具有正當性呢？Chomanski （2019）認為創造 AI 僕人是不道德的，依他的觀點，若 AI 能如設計者所宣稱的達到人類智慧水準，則它應該被視同與人一樣具有權利和責任。同時，即使程式設計的技術能達到，他認為想要設計出可靠的AI僕人就是一種操控慾的表現，缺乏道德直覺，令人反感。Bryson（2018）則討論 AI 系統是否能或是應該被賦予道德主體（moral agent）或道德受體（moral patient）的角色。他認為在技術上，或許可以創造出具有道德行為能力的 AI 系統，但是我們社會也可能不斷地重建道德體系。該文涉及倫理學的理論，Bryson 將 AI 在社會的地位這類議題歸類於規範倫理學，而非敘述倫理學的範疇，因此認為需要著墨的應該是設計 AI 時應遵守的規範。值得一提的是，此文將強 AI 列為關鍵詞，但文中未針對強 AI 提出論述，僅從結論看出作者的立場是認為不論是道德主體或受體的強 AI 都應避免。

道德主體是指該個體具備道德思考、自由意志而能採取道德行為，主張動物權利者認為動物是道德受體，牠們具有感受、意識或其他心智能力，但無法運用道德原則回應外在環境後採取行動，仍應享有生存權利（柯志明，2011）。De Winter（2018）以動物行為研究已知的物種個體行為差異為基礎，指出 AGI 可能發展出「性格」，這是技術專家們忽略的問題。動物有生殖、遺傳和適應行為，這些行為能否套用在 AI 上，使其達到真正的 AGI 物種，該文中作者呈現幾個對照表，建議在技術上應可朝向多種行為變異的 AGI 發展，甚至提出應為未來可能需要的人與 AGI 互動倫理準則做好準備。

一些 AI 研究者致力於發展具有道德推理判斷和決策的 AI 系統，使其成為道德主體或道德行為者（例如：Anderson & Anderson, 2011；Wallach & Allen, 2009），以避免未來假想的 AGI 機器超越人類智慧的科技奇點（technological singularity）之發生。擁護 AI 科技發展者提出機器倫理（machine ethics）的概念，是從系統設計端給予機器倫理原則或程序，讓機器得以發現並解決其可能遇到的倫理困境的方法，也就是讓機器自己能夠透過倫理決策，以符合道德的負責方式進行運作。哲學學者 Boyles

（2018）對此概念提出質疑，他認為機器倫理的概念與屬於應用倫理學領域的科技倫理架構有很大的不同，科技倫理是規範設計和使用科技的人，而機器倫理則視機器為道德主體。若未來 AGI 真的實現，則可發展出它們自己的倫理道德標準，到時將會挑戰人類對於現實、生命的道德觀。

另一篇文章討論 AGI 對人類造成什麼影響（Livingston & Risse, 2019），以下的問句耐人尋味，當 AI 機器和人一起存在這個世界上，區分他們是否有意義？機器欠人類什麼，人類又欠機器什麼？倫理的發展意義又是什麼？這兩位作者專精於數位科技政策的研究，他們知道目前 AGI 發展成為超級智慧的可行性仍受到質疑，但有必要思考 AGI 對人權可能衍生出的倫理道德問題。這篇文章最終提及機器也可能具有道德地位，而 AGI 機器將有利於人類還是侵害人權，尚未可知。

二、開發具有倫理思考的自主武器？

Courier 專刊討論 AI 帶給人類社會的威脅之一即為致命自主武器系統（lethal autonomous weapon system, LAWS）的發展，為 AI 技術在國土保安和軍事裝備領域上的應用。「自主」這個詞的概念對 AI 研究人員而言是用來隱喻各種不同類型的計算行為，但該詞富有多重意涵，常被解讀為代理者（Johnson & Verdicchio, 2019），可能成為超越人類控制的殺人武器。

一群分別來自義大利、美國和加拿大的學者 Umbrello、Torres 與 De Bellis（2019) 特別關注具有倫理規範的致命自主武器的開發。他們首先針對是否有全球共識的倫理道德規範以及參與戰爭人員心理和情緒上的影響提出論證，並以國際人道主義和相關法規中的「比例原則」（principle of proportionality），即戰爭時軍事人員和平民百姓的傷害比例難以主觀認定及估計，若能透過 AI 演算將可較精準的推估。因此，他們贊成應發展有道德的致命自主武器（moral LAW），其他的不具倫理道德判斷的致命自主武器都應該被禁止。他們的論點是自主武器的精確性和有效性高，若能根據戰爭法（LoW）和交戰守則（RoE）制訂出道德規範和管制措施，則可降低濫殺誤傷的可能性。既然各先進國家已經在發展致命自主武器，且科技發展原本就與國家社會的價值觀緊密結合，基於國際共識的法律，設計出具有倫理道德的殺人機器人，反而可造

福大多數的人。然而，此觀點明顯與印度哲學學者 Chakraborty（2018）的論題：「機器人可以有道德嗎？」有很大的歧見。Umbrello 等人認為自主機器具有比人類強的理性，或不受非理性成分的干擾，而道德是理性行為的表現。Chakraborty 則從倫理學的角度提出辯證，他認為理性或道德思考是有意識的人類和動物才能表現，此處牽涉到 AI 當中的智能這個字，而思考與智能有關。具有智能的生物是不會在沒有目的下完成任務，也不會在毫無動機和意圖下獲取或理解知識。AI 機器人不具有生物性歷程，「無法發展出情感、良心、理性或自我知識（self knowledge）來傳達其道德判斷的任何影響（p.53）」。相反的，來自歐洲國家同樣是哲學學者的 Lara 與其醫學教育的同事 Deckers 則認為 AI 可用來增強個人和群體的道德，他們用蘇格拉底的助手為比喻，主張透過人機互動關係可提高道德判斷的決策品質（Lara & Deckers, 2019）。

Solovyeva 與 Hynek（2018）的文章則探討了自主武器系統的六個難題，包括：自主武器系統性能的可預測與不可預測性、殺戮決定的非人性化、將敵方戰鬥與非戰鬥人員去除其人格、協調操作中的人機關係、戰略考量以及自主武器系統運作的法律範疇。這兩位來自捷克的學者的專長為政治和安全研究領域，以這六個難題為主軸，從是非優劣的正反兩方觀點，分析自主武器系統研發所引發的倫理、法律、政治、戰略和科學的對話。他們希望藉由提出此類分析架構，喚起人們對自主武器系統的認識，並期許後續的研究能針對上述難題進行診斷和尋求解方。此外，他們強調在任何 AI 和自主武器系統中，仍不可缺乏人類的監控，人機互動關係的建立是重要的環節。一群 IEEE 會員 Brutzman 等人（2018）研究無人駕駛機具的能力和局限性，以及人類如何下達命令責任和權威的真實示例，因而支持道德機器人的可行性。他們認為，即使是執行致命性的任務，只要定義清楚的指令和人機協作關係，自主機器人系統是能夠完成符合倫理道德規範的使命。他們不僅提出技術發展上的概念流程圖，也透過模擬試驗確認無人駕駛的自主載具（包含自主致命武器）有能力解釋人類所定義的戰略目標或戰術任務。該研究提供了以倫理道德為基礎且在人類監督下，運用無人系統執行遠端任務的基礎架構。

另一方面，Gill（2019）則擔心在未來 20 年會因為新的 AI 自主武器霸權產生國際衝突，各個 AI 武器強權國家應該藉由多邊會談建立信任機制，以制定武器平衡的規範。他認為致命的自主武器技術用於戰爭，可能會產生無差別式的攻擊，將會挑戰

國際人道法的原則，此與 Umbrello 等人（2019）有很大的歧見。Gill 認為全面武器的應用程式如何來做區分和預防等措施，都需要有一定程度的人類反思和控制，因此在維護國家安全上，迫切需要各方面的管理與規範，以降低國際戰爭的風險，並維持世界和平與安全。這位印度大使在聯合國擔任審查致命自主武器系統新興技術的組長，曾於 2018 年 4 月 9 日，在瑞士日內瓦召開的政府專家會議上鄭重宣布：「各國必須對本國部隊在武裝衝突中致人死亡的行為負責」（UNESCO, 2018, p.27），代表聯合國對於致命自主武器的開發是持保留態度。而有限度的自主武器可否透過設計有道德的機器和人機協作關係達成，目前雖有前端技術概念（Brutzman et al., 2018)，但國際間在治理和立法規範上尚未有共識。

三、AI 在醫療上能成為人類的「助手」？

「未來的機器人也不過是我們忠實的助手而已。」（UNESCO, 2018, p.11）Courier 專刊的幾篇文章皆如此信心喊話，其中一篇更報導 AI 仿生手的技術如何造福肢體殘疾者。自 1980 年代之後，AI 系統在醫療診斷和決策上已逐漸獲得廣泛的應用，但也引發許多尚待解決的道德、隱私和生物倫理等問題。

與仿生手有關的案例，如同 Droste 等人（2018）文章所提及的智慧植入物或假體，結合人體的改造，在優化醫療方法上獲得長足的進步。如 Rybarczyk 等人（2018）以植入髖關節假體的醫療處置為例，透過網路平臺遠程監控術後患者運動復健情形，顯示良好的成效，但他們未針對可能發生的個資外漏和駭客入侵的問題加以著墨，在提出此類網路醫療平臺使用介面的開發建議時，也忽略解決上述問題的檢核機制或相關的技術。然而，這些智慧植入物的設計是否符合我們的法律制度？是否應將倫理價值觀納入這些技術物當中？又，如果機器是從人身上複製的，是否必須具有法律主觀性？這些問題是德國學者 Droste 等人為文探討的重點。他們提出醫療設備應該要符合高標準的健康與安全保障，具有自主學習能力的 AI 假體發生問題時，醫療保險的責任歸屬皆有法令規範未竟之處。文章中提到植入物與人腦有對接埠，而 AI 的類神經網路技術如同黑盒子般，植入的機器若功能失常而影響到人的行為或決策錯誤時，將衍生出責任歸屬的問題；又若 AI 假體常與物聯網（IoT）有連結，當其介面

受到駭客入侵，製造假訊息或是盜取訊息，甚至毀壞植入的假體，可能引發的隱私和安全問題。這些問題都直接影響到這類醫療產品的製造過程與開發的細節，因此需要有詳細的倫理規範以及責任歸屬的法律限制。

AI 在醫療上是否能成為人類的助手？McDougall（2019）同意如果醫生和患者之間在決定醫療方案時，能由價值判斷有彈性（value-flexibility）的 AI 系統協助共同做決定，將可提升患者自主權，使醫治過程更有效且更符合倫理的方式。然而，他隨後批評 IBM 公司設計的「Watson 腫瘤科」（Watson for Oncology），認為該系統設定的治療選擇是以「壽命最大化」為價值目標，並非立基於個別病患的價值和意願，對於醫病間的共享決策過程反而是有威脅的。McDougall 認為 Watson 系統透過大數據和演算結果，提供醫生與病患多種治療選項和決策上所需的資訊，但是病患個人的價值觀並未被納入評估，與共同決策的概念有所違背。Di Nucci（2019）對 Watson 腫瘤科案例卻有不同的看法，他指出 McDougall 的論點存在三個問題：（1）將 AI 與機器學習混為一談，Watson 是機器學習演算法，不能稱為 AI；（2）低估了機器學習倚賴大數據可針對個人做出醫療建議的潛力；（3）無法區辨在醫療保健中以實證為基礎的建議和個人做決定的過程。Di Nucci 試圖釐清此 Waston 系統是基於大數據和演算法提供醫療上的實證建議，若將病患個人意志納入，則非醫療實證，也對共享決策不利。他其實是批評科技末世論（techno-apocalypticism）的觀點，認為人們不應恐懼 AI 演算法在醫療上的應用。

關於 AI 與機器學習在醫療上的應用，公平性（fairness）是醫師們最在意的問題之一。Rajkomar 與其醫師同僚特別指出機器學習應用於醫療保健時，可能出現偏見的四種狀況：（1）模型設計中、（2）訓練數據中、（3）與臨床醫生互動時，以及（4）與患者的互動（Rajkomar et al,. 2018）。放射科的醫師們 Balthazar 等人（2018）主張制定「患者權利法」以保障病患的隱私、數據資料所有權和知情同意權。Kirkpatrick 與 Pearlman（2019）則討論心臟超音波影像應用在心血管疾病醫療上所面臨倫理挑戰，他們認為 AI 未來不可能取代心臟超音波檢驗師的地位，只有醫事從業人員才能展現出利他主義、正直、責任感和尊重等專業素養。相反的，Fiske、Henningsen 與 Buyx （2019）相當支持 AI 機器人應用於精神科與心理治療上，其好處是可嘗試新的治療方式、能觸及某些難以接觸的對象、獲得患者更好的反應並減少醫生治療的時

間，同時他們也討論了倫理道德與法律規範之間的落差，以及風險評估、監控機制、尊重病患自主權、演算法的透明度以及療效的長期追蹤等倫理相關的議題。

Weber（2018）探討醫療機器人衍生出的醫學倫理問題，他將醫療機器人細分為七種，包括：假體、照護、手術、影像和導航、決策、仿生、自動投藥等用途，亦論及醫療機器人仍有許多未解決的問題，如機器的自主性、同理心和情緒相關的觸覺經驗、動機和信任（包含數據隱私）以及對社會和經濟的影響（包含工作取代和去人性化）等人際關係的議題。然而，對於這些倫理議題，Weber 並未提出積極性的解方或建議。專長於醫學法律與倫理的 Schönberger（2019）則分析該領域的重要文件，他檢視了約 300 份政府和非政府組織所發佈的 AI 和醫療保健相關的公共政策檔案，以及相關學術和科學性論文，觀察到 AI 醫療應用的問題多圍繞在以下的倫理與法律相關內容，包括：（1）公平與歧視，（2）自主性和訊息/取用的權利，以及（3）道德責任與義務。文中探討了一些例子，他認為目前一些尚未成熟的硬性法律可能扼殺許多有利的創新，所以應該要針對個別部門訂定法規，以回應人們對 AI 系統決策能力的擔憂，如：偏見、透明度和模擬的真實性等；同時，也應該激勵負責任的 AI 技術的研發，讓這些技術能監控上述問題是否在醫療保健領域發生，且能達到公平、尊重個人自主性和資訊權，在道德責任和義務上，Schönberger 也建議「人在環內」（human in the loop）的作法，下一節將進一步討論此概念。

四、AI 衝擊勞動權益關係？

「大數據驅動的 AI 科技正在推動第四次工業革命」（UNESCO, 2018, p.22），將撼動全球的社會經濟。第一次工業革命的機器化取代人力，所產生的社會動盪仍殷鑑不遠，而 AI 科技應用於商業自動化造成勞動市場結構性的演變，目前正在許多國家上演中。

Wright 與 Schultz（2018）的文章，開頭即以美國 Amazon、韓國現代汽車以及在羅馬街頭實際發生勞工抗議 AI 科技產品研發的各國案例，討論商業自動化造成文化與倫理的影響。他們定義商業自動化為以機械或電子方式代替人工操作或控制商務流程的技術、方法或系統，對於權益關係人（stakeholders）包括：勞工、企業、政府、

消費者和社會，有不同程度的影響。商業自動化可降低成本和生產時間，提高產量、安全性和品質，企業方必然選擇朝自動化方向發展，因而改變勞動市場的結構。從消費者的角度來說，自動化高效率可能使商品價格降低，有利於增加購買力。國內雜誌媒體近期也有許多報導談論哪些職業未來可能被 AI 取代，主管和勞工的薪資水準也可能有所改變（如：鍾張涵，2019）。Wright 與 Schultz 繼而從社會契約理論提出整合性的倫理架構，以考量不同權益關係人的需求，以及減少 AI 對勞工階層帶來的衝擊。

相對於擔憂大型企業採用商業自動化對勞工產生衝擊，來自日本和荷蘭的 Kamishima、Gremmen 與 Akizawa（2018）在一篇管理哲學的期刊論文中，認為權益關係人應共同合作，為 AI 機器人的創業家制訂一個行動許可範圍。他們提出制訂創業方案的過程即是多元學科的策略，在講究效率過程中強化以人類能力為核心的管理目標，給予 AI 機器人創業發展的空間。此外，Mendling 等人（2018）也提出機器學習、機器人流程自動化和區塊鏈等 AI 技術，藉由自動化效應、資訊效應和轉型效應等三種效應，改變業務流程管理（business process management, BPM）的生命週期。BPM 生命週期包括：發現、分析、再設計、實施和監管等步驟循環進行，在這個循環流程中，前述的 AI 相關技術不論在個別任務層次或業務協調層次，可減少人為因素介入，使業務流程更公平，更不容易受到貪腐的影響。文中最後也提及過去兩個世紀以來，自動化和科技進步從未停歇過，而人力勞動的需求也不會因此斷絕，所以因應AI與商業自動化的影響，應該思考的是工作設計（job design）、顧客經驗，以及人們對新科技的接受度。另外，也建議針對區塊鏈和數位貨幣等技術應有管制規定，因為區塊鏈企業主期待有明確的管制，如此才可確定企業的合法性和稅收。

而對權益關係中的消費者方面，Wright 與 Schultz（2018）提到 AI 機器進入人類生活，會降低人們對社會的歸屬感和幸福感。Mendling 等人（2018）則提及引入 AI 技術的資訊科技產品，讓人們感到更快樂，對生活更加滿意。Wirtz 等人（2018）在服務管理期刊的文章，也推薦服務型機器人的設計與應用可提高服務業的品質。這幾篇文章所探討的 AI 技術各有不同，對於在商業上的影響，不論是持謹慎保守或樂觀開放的態度，皆認為這類主題的研究應該是跨學科領域合作，也都提到社會公平性和勞雇間權益關係的道德考量議題。

伍、AI 倫理治理對策

延續前述的議題探討，本段內容主要選介文獻中針對倫理規範和治理框架所提出的觀點，分成四個主題，從較為上位的系統研發之規範準則制訂，到較為具體的治理技術之探討。

一、自主與智慧系統研發規範

對應於前一節的強 AI 和自主武器是否應該開發的議題，毋庸置疑的是，世界上許多先進國家已投入智慧自主系統的開發，目前已有許多突破性的成就，從事此領域的研究和實務工作者往往更期待其產出能獲得公眾的肯定，讓研發工作更具正當性。

電機電子工程師學會（IEEE）是 AI 研發的主要學術團體，如前所述，該學會於 2016 年即倡議訂定 AI 設計的倫理準則，學會會員們相信倫理準則的訂定可彰顯該組織的價值信念，並建立與客戶和權益關係人的信任關係。來自澳洲、美國與法國的 IEEE 會員 Adamson、Havens 與 Chatila（2019）撰文闡述 IEEE 在倫理治理的原則和策略方案上扮演領航者的角色，列出 IEEE 於 2018 年發起的 52 項活動、計畫、團體或教育方案。IEEE 下屬的幾個組織，包括：科技的社會影響（SSIT）、機器人與自動化、醫學與生物工程學等學會，這些組織的研究主旨都關注於倫理價值觀。IEEE 提出智慧自主系統倫理的全球倡議，其中《以倫理為基準的設計》（Ethically Aligned Design）於 2019 年發佈最終版本（IEEE Global Initiative, 2019），該文件旨在教育參與智慧自主系統開發的權益關係人能優先考量倫理和人類福祉，也是後續 IEEE P7000 系列許可標準的依據；而 P7000 計畫整理 1300 多份全球 IEEE 的標準，多數的標準與技術的互通性、安全性和貿易便利化有關，而這項計畫則著重技術和道德考量之間的問題。

來自英國的學者 Winfield 和 Jirotka（2018）發表論文探討機器人 AI 系統的倫理治理，此文刊登於英國皇家學會期刊《自然科學會報 A》（Philosophical Transactions of the Royal Society）物理科學類。他們彙整了包括 IEEE、阿西洛馬以及橫跨歐盟、美國、英國和日本等政府單位提出的倫理原則相關文件共十件，在文中感嘆倫理準則

氾濫，且倫理治理方面鮮少有良好的實踐，因此試圖從倫理學—標準—管制的路徑圖，提出良好的智慧自主系統倫理治理的五大支柱。第一支柱是發佈倫理道德的行為準則（code of conduct），第二是讓所有人都接受關於研究與創新倫理的訓練，第三是實踐負責任的創新，讓更廣泛的權益關係者參與科技創新的治理，第四是讓倫理治理及其管制過程公開透明，最後是任何組織都要重視倫理治理的必要性，甚至將倫理治理比喻為公司遺傳密碼 DNA 的一部份，如此才能讓智慧自主系統獲得公眾的信任。發表在同一本期刊的另一篇文章，來自英國牛津大學圖靈學院的學者 Floridi（2018）著重在數位資料相關的科技創新治理的問題，討論了軟性倫理（soft ethics）和硬性倫理（hard ethics）的差別和彼此間的關係。硬性倫理是在制定法律時用於判斷道德上的權利、義務和責任關係，軟性倫理涵蓋與硬性倫理相同的規範基礎，但更廣泛考量人權、服從性和可行性會隨著時間演變，因而從自我規範的角度探討在現行法規之外應該做和不應該做的事。他用一張圖描述倫理、管制與治理三者之間的交互關係，指出硬性倫理透過社會接受度或偏好度影響管制與治理手段，管制則因遵守法規而限制或達成治理的成效，同時也影響軟性倫理。Floridi 主張應該以軟性倫理為架構應用於科學、工程、科技與創新的領域中，他提及公眾對 AI 科技的接受度和使用度的產生必須是在科技效益對公眾是有意義的，且潛在的風險可以預防或最小化的狀況下，因此提出了前瞻分析週期環和倫理影響評估的模式。

美國學者 Kroll（2018）從技術面談倫理治理的問題，他認為電腦系統不是純粹中性的工具，而是其社會系統背景下的產物，不應以「黑盒子」的比喻來搪塞其應有的倫理道德管制，因為演算法本身是從根本上可以理解的技術。因此，對演算法有效的治理方式，須嚴格要求科學與工程的系統設計、執行和評估，使系統的信賴度具有可驗證性。同時，系統的假設、選擇和正確性的決定必須是可公開被檢驗的，因為問責（accountability）、透明度（transparency）與公平性（fairness）是此類智慧自主系統及其演算法倫理治理的重要原則。

探討不同國家組織對於 AI 倫理規範的願景和策略有何異同，Cath 等人（2018）分析比較了美國白宮、歐洲議會和英國下議院分別針對社會如何因應 AI 廣泛應用所提出的報告。這些報告並非倫理規範的條文，卻代表政府官方對於 AI 科技治理的意象。這群作者包含前述的 Floridi 等英國牛津大學的學者，根據他們的評估，英國的報

告試圖指出 AI 與機器人科技的潛在價值與功能，同時要檢視此科技可能需要預防、減緩和治理的預期問題和不良後果。相對於美國對 AI 科技管制措施傾向於「百花齊放」（根據作者所下的標題 letting a thousand flowers bloom）鼓勵 AI 的研發，英國的觀點較像是「靜觀其變」（keep calm and commission on），而歐盟的報告則呈現出對 AI 機器人取代人力的擔憂，且對可能的風險採取硬性和軟性的法律規範，因而針對特定科技系統開發（如：自駕車、無人機和照護機器人）有倫理治理規則的制訂，如強制險和機器人註冊方案。三份政府文件的比較結果，對社會有正向影響的 AI 科技是各國共同的期望，透明度和問責是共通的治理指標，但治理目標上有一定程度的差異。AI 科技政策和倫理治理模式的跨國比較，較缺乏亞洲的中國和日本等國資料，是此次文獻蒐集分析力有未逮之處。

二、倫理價值觀嵌入 AI 系統

倫理治理的需求受到國際上 AI 科技發展大國的重視，而 AI 科技的範疇亦相當廣泛，與其由政府組織制訂法律規範，不如由開發端自發性的建置具有倫理道德規範的 AI 系統。然而，倫理價值觀如何嵌入 AI 系統？是否衍生責任歸屬的問題？以自駕車和無人機為例，運用類神經網絡的機器學習技術所製造的 AI 系統，實務上出現的狀況是：操作機器的人員已經不再能夠預測機器行為，如此是操作人員還是 AI 系統應被追究道德責任或承擔後果呢？若將 AI 設計為道德行動主體者具有道德意識，如同 Levy（2009）的疑問：「我們怎麼知道一個所謂的『人工意識』機器人是否真的有意識，而不是僅僅表現為有意識的樣子呢？」（p.211, 引自 Hoffman & Hahn, 2018）。

Hoffmann 和 Hahn（2018）認為從道德主體與道德受體的角度(前段議題中曾討論過)，談論倫理的 AI 系統（ethical AI system）具有的性質和道德地位，所耗費的成本太高，於是他們嘗試從哲學的角度分析倫理學對法規前景展望的影響，從而提出較為實際的政策建議。其文章的標題以去中心化的倫理為名，抱持著相對主義的哲學觀。他們認為在道德實踐上，AI 系統可以被視為具有道德意義和感知力，換言之，要定義人類與 AI 之間的道德互動途徑，並非透過外部的、絕對的框架去釐清責任和道德主體之間的概念問題即可，而是要透過社群內部共同關注哪些關鍵因素是具有重要道德

地位而做決定。他們建議採用某些支持決策的基準（benchmarks）測試機器的道德狀況，基準指的是讓我們定義 AI 需要滿足的最低能力要求，而不是設定嚴格標準規定不適用的所有條件。此外，他們也要求 AI 系統不能與人類混淆，亦即擬人化的 AI 設計應該要避免，讓 AI 與人類的互動，定位在我們對機器的理解程度，而不是去討論機器模擬人類時的道德狀況，如此才不會落入道德主體和受體的爭議。最後，他們對政策的建議，包括強化非政府組織（如 AlgorithmWatch）出席相關的政府委員會，在塑造公眾對話上發揮重要的功能，以及強制要求開發者提供 AI 影響分數，讓人們能夠公開評估 AI 的應用。

道德兩難困境的案例常用在探討 AI 倫理治理議題上，de Swarte、Boufous 與 Escalle（2019）嘗試探索這幾個問題：AI 比人類智能更符合倫理嗎？AI 比人更尊重人的價值觀嗎？他們試圖探討是否可以功利主義方法促進倫理，此方法是指在一系列可能性的情況下，選擇導致行動的解決方案，此解決方案能最大程度地發揮內在良善或達到淨愉悅（net pleasure）。他們提及 Ethicaa，這個研究方案是在 AI 中植入倫理道德原則，討論人工行動主體者（如：戰鬥無人機和陪伴機器人這兩種極端的自主 AI 系統）是否有能力做出倫理決定。作者的分析認為軍用無人機在戰鬥中比人類駕駛較不會受到情勢上的壓力，因此較能實行道德的行為並尊重人的價值，特別是當演算法中已經納入發射可逆性（shot reversibility）的概念。相對的，陪伴機器人在設計上似乎是幫助改善老年人的健康狀況，雖然目的是良善的，程式設計也不會傷害人類，但是讓機器人變成受照顧者每日生活不可缺少的模式，甚至因必須聽從機器人的命令而失去人類的尊嚴。相較之下，陪伴機器人對人類的社會、心理和身體尊嚴有潛在的危害，若是程式設計無法因應新形式的歧視、偽造、假訊息、誤判和社會規範時，產生的倫理和法律的問題更大。如同先前介紹過 Weber（2018）針對醫療機器人的論點，這類機器人就像是「參與具有道德後果的行為，但由於缺乏自主指導的意圖而不能承擔道德責任……它們充當著沒有道德責任的道德行動主體者」（p.605）。因此，de Swarte 等人建議陪伴機器人應該被設計為護理人員的助手，照著人類的指令照顧好病人，而不能取代護理人員的角色。整體而言，他們認為功利主義方法特別是採用智能主體為本的理論（agent-based theory），賦予自主 AI 系統有能力根據所涉及的倫理原則區分出最理想的選擇，此法可被視為 AI 倫理治理的解決方案。

有關 AI 可能侵害人類尊嚴的觀點，Kanuck（2019）在文章中也提到，AI 伴隨著機器學習演算法和專家系統，可能取代許多人類在做的事情，如：醫療診斷甚至是運動賽事的報導，讓人們不免質疑人類還有什麼剩餘的優勢，又如何維護人性和人道行為的觀念。Kanuck 認為幽默和愛是人類獨有的行為特質，而倫理標準中重要的概念如「正義」和「公平」這類複雜的目標，屬於高層次認知。人類在價值判斷和做決定的歷程中，往往會注意到超越特定情境和時間之外的附屬訊息，能夠歸納經驗中大量不相關的資訊，用來做出或改變預期的結果。他以無人車和自主致命武器為討論的案例，認為人之所以和處理資料的 AI 機器有區別，乃是因為人類能夠處理不協調資訊（incongruous information），以及能觀察到細微差異的脈絡情境，而這些是機器學習演算法無法辦到的。他最後提出兩個選項：（1）我們是否要對 AI 施加限制以利於維持人為控制；（2）我們是否要在 AI 機器中，複製人的道德推理和橫向思維。選擇不同的選項和答案，將對於人類與 AI 的互動，以及我們人與人之間互動的方式，都會產生深遠影響。

三、偵測 AI 歧視、偏見和犯罪

人類社會對於歧視、偏見和犯罪，皆制訂了判斷準則，是依據人與人互動經驗所累積建立的倫理關係。而當 AI 機器成為代理人，有可能複製人類的意志，或透過學習（超級智能），發生對人的歧視、偏見和犯罪行為時，所需的防範處置即為人機互動倫理的重要議題。

Howard 與 Borenstein（2018）指出偏見是如何被置入於AI和機器人系統中，特別著重在偏見是如何影響維和機器人（robot peacekeeper）、自駕車和醫療機器人等的功能。偏見會干擾決策過程，此處強調的「隱性偏見」（implicit bias）是指相對無意識和自發性的偏頗判斷及其呈現出的社會行為，如：刻板印象，是社會文化長期累積很難根除的價值觀，甚至發展成為在高風險情境下增強決策過程的保護機制。除了前述三項常被討論的 AI 系統，文中還提及幾個 AI 演算法的偏見案例，包括：人臉辨識系統對於不同膚色人種影像的錯誤標記、語音辨識能力獨厚男性的聲音、搜尋引擎中的資訊偏差等。套用前段引述 Kroll（2018）的觀點，演算法不是價值中立的，系統設計

者有責任確保其透明度和公平性。Howard 與 Borenstein 也同樣敦促 AI 專業者自律，拋開個人和機構本身的偏見，透過參與式設計方法，包容社群或公眾的意見和共同監督，以避免偏見所造成的社會不平等的問題。

這些學者主張 AI 演算法是反映出社會權力結構、眾人期望、信念和價值觀，因此需要對設計端有所約束。然而，科技政策的制訂和執行往往處於兩難狀態，應該讓多數人獲得方便快速的利益，還是要顧及所有人且不犧牲少數人的權益？即使演算法的設計和模擬能盡可能達到公平的原則，但仍舊會因為數據的採樣不足或偏差，使得機器學習時訓練方式和優化過程，發生偏見或歧視的問題。Turner Lee（2018）認為演算法是有助於自動化決策，但是偏見是這些演算法的副產品，會對弱勢族群造成危害。其研究是分析在美國發生的案例，因為大數據排除某些族群，或大數據分析的不當應用，而追蹤線上使用者的活動和行為，使得強烈倚賴大數據的演算法和機器學習產生種族歧視的結果。他也提及隱性偏見會出現在機器學習的複雜計算中，這類案例發生在語言翻譯、犯罪預測、以及從名字辨別性別等由 AI 科技應用導致偏差的結果。他將這些問題歸因於高科技產業中缺乏種族、性別和文化的多樣性，並且提出 AI 演算法如何避免種族歧視，方法一是統計上的均等性（statistical parity），即機器學習的訓練資料來源必須來自不同族群且具有相等比例，方法二是在條件上具有統計均等性，即針對一組可能的風險因素控制在相等比例的受試者中進行，方法三則是預測的平等性，必須假定決策正確率和錯誤率在各個種族群體中均相同。該篇文章最後建議，就 AI 科技創新領域而言，政策制定者和技術專家應當遵循「無偏見」的原則，從公共政策和律法上提供有色人種就業、居住和信貸保障，以及更多樣化的工作機會和場所，如此才能使演算法成為通往機會、平等和效率的橋樑。

另一方面，英國圖靈學院 Floridi 等人（2018）組成的 AI4People 團隊，則致力於以確保人們信任、服務於公共利益並增強共同的社會責任的方式開發 AI 技術，利用 AI 糾正過去的錯誤和消除不公平與歧視。King、Aggarwal、Taddeo 與 Floridi（2019）系統性地分析跨領域文獻，從對社交媒體用戶的自動化詐欺到 AI 驅動操縱模擬市場的案例中，預見 AI 犯罪的威脅。其分析的文獻中，不乏有研究建議用 AI 偵測 AI 犯罪，如同以其人之道還治其人之身，例如 IBM 設計將認知技術導入資安監控維運中心的平台（IBM Cognitive SOC），其應用機器學習演算法提取網路資訊如網誌、文章、

報告內容，偵測識別、降低並回應其中具有安全性威脅的訊息。然而，King 等人提出警訊，若是過度依賴 AI 仍可能適得其反，後續應針對犯罪行為的個人心理與社會因素做深入的研究。

一群日本學者 Ema 等人（2019）的研究是分析一篇論文探討同人小說（fan fiction）寫作中的價值衝突，從線上同人小說文本中，提取和過濾有關性的表達方式，藉此說明隱私的概念，並作為 AI 倫理治理的經驗教訓。此文所描述的倫理爭議，是有一群學者應用 AI 技術從事涉及隱私和敏感議題的研究，在一場日本的電子工程學研討會上發表後所延伸出來的議題。該研究運用 AI 技術辨識文本內容，並宣稱其研究目的是為了偵測淫穢的語詞，以移除對青少年有害的資訊，然而該研究所使用的資料來源是可識別的，即違反歐盟的《一般資料保護規則》（General Data Protection Regulation, GDPR）和日本總務省訂定的 AI 倫理準則，因此在研討會發表後引發網路社群的撻伐。此案例中，AI 技術原本似乎是使用在良善的用途上，但是當公共和私人資訊的界限不十分明確，且當事人並不知情或同意其內容被公開使用時，研究人員藉由機器看似無意識地使用這些資訊，卻實際上侵犯了法律地位薄弱的社群。這篇文章最後建議 AI 倫理治理，不僅止於設定 AI 設計的目的和產品應用需遵循倫理原則，更是需要跨領域合作共同研究倫理範疇的課題。

四、人在環內（Human-in-the-loop）的干預機制

關於 AI 倫理治理的文獻中，Human-in-the-loop（人在環內，或譯為：人在迴路、人機迴圈、人機共生）高頻地出現在七篇論文中（如：Borenstein, Herkert, & Miller, 2019；Kanuck, 2019；Rahwan, 2018；Green, 2018；Schönberger, 2019；Shank, DeSanti, & Maninger, 2019；Zanzotto, 2019），該詞的主要意涵是機器系統有從收集資料、判讀情勢到做出決策的迴圈，而人必須參與其中。這個概念在 20 世紀末期已被提出（Secretary of Defense for Acquisition Technology, 1998），在 AI 治理的領域中，主要的意涵即是系統無法提供問題答案，需要人為干預，人類利用系統所提供的資訊做進一步的判讀，以人機合作的方式，結合可變更運算參數，最終能精準有效地修正 AI 演算法所做的決策。

義大利學者 Zanzotto（2019）提出人在環內 AI（HitAI）是一種負責任 AI 的典範，目的在提供知識生產者正確的價值觀；任何 AI 系統都應該有人的參與。所以，Zanzotto 認為 AI 自主學習必須是具有解釋性的，程式設計本身就應該將人置於迴圈中，純粹讓機器自主學習是一個不適當的模型。HitAI 系統強調，在特定情況下必然使用了「某人的知識」，因此原始知識生產者應享有信譽和薪資收入。Zanzotto 提出 HitAI 系統的構思，主要是回應就業市場可能被大財閥壟斷的問題，但也不諱言在實現時會面臨兩個重大的挑戰：第一個是如何說服公司與數據生產者分享利益，第二是如何建立可信賴的 AI 系統知識生命週期，以及如何管理知識所有權等基礎結構和維運的挑戰。

更進一步，美國麻省理工學院的學者 Rahwan（2018）提出「社會在環內」（society-in-the-loop）的概念框架，即「社會契約」與「人在環內」的組合，讓公眾社會納入機器訓練與決策的迴圈中，這樣才能展現出 AI 更可靠的代表人類，並且包容更多元的價值觀。他認為透過許多專家學者以及決策人員的參與，才能確保AI演算法公開、透明、公平且負責，符合前述的倫理治理的主流觀點。人在環內的系統雖以強調人類的判斷（包括目標、限制、期望和知識等）指導監督自動系統的數據、感測器、演算法、統計模式和使用功能等，但顯然只把控制權交給某個或某群人類專家，缺乏以社會為整體的角色。同時，Rahwan 與其同事們在《自然》（Nature）期刊發表一項實驗（Awad et al., 2018），稱為道德機器實驗（moral machine experiment），他們以自駕車為實驗情境，設計了道德兩難判斷的遊戲，為時兩年，收集到來自 233 國家共約 4 千萬筆資料，其中有將近五十萬份資料有人口學統計資料作為跨地域分析的基礎。基於這個研究結果，Rahwan 主張機器學習演算法應將社會價值觀、喜好和期望納入設計、監督和管控迴圈中。在 Rahwan（2018）這篇文章中提出如何達到「社會在環內」的目標，其一是運用設計、眾包（crowdsourcing）和情感分析具體呈現利益關係者和大眾的價值觀，其二是以契約主義的觀點，協調社會群眾偏好並達到資源公平分配，最後則是合規性（compliance），即運用演算法觀察和監督演算法，他引用 Etzioni 與 Etzioni（2016）提出的監督程式演算法，認為該演算法可用於監視、審計和追究 AI 程式操作的責任。他強調決策迴圈中的每個環節都需要清楚定義，並引入人類利益協商結果，也就是納入社會契約於 AI 系統編程、除錯和監督的流程中。

綜合前述有關人或社會與機器協作於決策迴圈的文獻，可再次檢視 Floridi 等人（2018）提出的 AI4People 的「良好 AI 社會」倫理架構，他們在文中約略提到「若將任務委託給 AI，人不在環內（in）或環上（on）（p.693）」，可能累積許多潛在的危害並影響廣泛，然而，他們同時也認為如果能將道德促進框架嵌入於 AI 系統，仍然能夠委託 AI 幫助人類擴大和加強共享的道德體系。他們主要的主張是從五個倫理原則，包括：仁慈、非惡意、自治、正義、可理解性等面向，對 AI 系統嵌入倫理框架提出討論，並蒐集七個官方或民間組織的文本，整理出二十條行動方針的建議。整體觀之，Floridi 等人是從較為「人在環上」的角度，主張做出符合倫理的好 AI，並從環後（post loop）的角度，確保使用 AI 能創造共享的利益，且不產生新的危害（如破壞現有的社會結構）。他們也試圖勸導大眾，不要因恐懼、無知、錯誤的關注或過度的反應，導致社會未充分利用 AI 技術的潛力，損失了機會成本。

陸、結語

本段綜整前述內容分析結果，並考量 AI 倫理相關議題的跨國共通性與本國在地應用，加入華人社會與文化的價值觀，綜合討論可能有助於我國政府研議 AI 倫理治理框架之啟示。

從文獻分析結果中，不難發現這個倫理議題的命題：「若未來 AGI 真的實現，會對人類造成什麼影響？」是多篇哲學論述的文章採用的基本假定。在 AI 科技的世界觀裡，人與 AGI 的思維方式或許可以解釋成兩個可以互相轉換的維度，就如同將適用於人類的類比符號——感覺、影像和文字，轉換成 AGI 能理解的數位碼一般。然若以「物質的因果關係」（Johnson & Verdicchio, 2019）解釋 AI 為非人類行為，則忽略了設計製造者欲賦予 AI 作為道德行動主體者之企圖，因此建議了第三方行動主體者的概念。同樣的觀點在 de Swarte 等人（2019）的行動主體者本位理論中出現，賦予 AI 判斷所涉及的倫理原則，使其自主做出選擇。想像當 AI 具有超級智能水準能獨立於人類做判定，也可能發展出如同人類一樣的「性格」（De Winter, 2018），必將改變人類與機器互動的關係，挑戰社會價值與道德觀。又若將倫理價值觀嵌入 AI 系統，或發展出具有道德意識的機器，其設定的倫理道德框架是放諸四海皆準的，還是因國

家、族群或文化而有不同？除了 Bryson（2018）提到社會的道德體系可能不斷地重建，尚未有研究考慮到跨國或跨文化間的差異。

有關發展倫理準則的倡議，Boddington（2017）曾評論「阿西洛馬人工智能原則」（Asilomar AI Principles）（Future of Life Institute, 2017）。該原則有 23 項條文，於 2017 年在美國阿西洛馬市的會議中簽署，強調 AI 發展的目的是為了造福人類，其中最後一條共好原則提到：「超級智慧」（與 AGI 概念相似）的發展應該僅能服務於廣泛認同的倫理理想以及全人類，而不是用於單一國家或組織的利益。Boddington 則認為倫理規範應更明確且積極，原則性的宣示可能流於空泛，而廣泛認同的倫理理想，也可能因難有明確的定義，使得少數聲音被忽略。

我國典籍中的「民胞物與」是一種具有睿智的胸懷，恰與「視萬物如芻狗」的不智相反，依此觀點，我們同意 Chomanski（2019）的主張，認為製造 AI 成為僕人是種不道德、有操控慾的表現。若 AI 成為道德主體或道德行為人時，也將如同討論動物應有生存權一般，AI 是否也應享有跟人類一樣的權利？而當 AI 犯罪時，法律上的究責對象又該如何界定？人們之所以恐懼 AI 科技發展，乃因 AI 彷彿是集眾人之智慧而成的一顆超級大腦，雖然可以為人類服務，也可能成為毀滅人類的武器，文獻中最具有爭議性的即是設計具有倫理的、且不受非理性干擾表現的致命自主武器（LAWS）。歐美科技大國和幾位 IEEE 學者傾向支持發展有道德的 LAWS，只要依據道德規範和管制措施設計自主機器人系統，則可降低戰爭中的人員傷亡，但印度和捷克等國的學者（Chakraborty, 2018；Solovyeva & Hynek, 2018）則持反對的態度，Gill （2019）更直指未來會因為 AI 自主武器霸權產生國際衝突。然而，目前幾個 AI 霸權國已積極開發前端技術，根據「阿西洛馬人工智能原則」中的共好原則（Future of Life Institute, 2017），這類超級智能應為全人類的福祉，而非個別國家或組織所擁有。當國際間在 LAWS 治理和立法規範上尚未有共識，學界立場也有極大的差異，此類科技的研發顯然對於人類社會是存在著相當大的威脅。

生老病死原是生物自然的限制，藉 AI 醫療機器人或者仿生假體來協助醫療照護，不啻為人類的一大福音，但也引發許多道德、隱私、責任歸屬的法律限制和生物倫理等問題。另以腦機介面（brain-computer interface）為例（Ienca, 2019；Miller, 2019），當透過此系統把人類的想法直接與互聯網連接，我們無法想像若有駭客入侵

到我們腦裡的想法時，我們是否還能稱為是智能主體？隨著肢體語言、面部表情的數據被收集與解碼，人類的行為表現可被分析，甚至複製於機器中。人類的思考或是反應速度早已不是機器人的對手，這種在商業殘酷競爭中的現實，已經使得利用 AI 帶動商業自動化、改變生產與勞動市場的結構、強化業務流程管理經濟效益，成為必然趨勢。然而，使用 AI 不論是持謹慎保守或樂觀開放的態度，仍應是跨學科領域合作，也都應關注社會公平性和勞雇間權益關係的議題。

有關 AI 倫理的治理，IEEE 這類的學術團體自發性的提出倫理治理準則，提出《以倫理為基準的設計》（Ethically Aligned Design）的倡議文件，還有 IEEE P7000 系列的許可標準，著重於技術的互通性、安全性和貿易便利化（Adamson, Havens, & Chatila, 2019）。AI 科技應用範圍已非常廣泛深入，為了防範 AI 複製人類的意志，又由機器自主學習而發生對人的歧視、偏見甚至犯罪行為時，必須要對設計端有所約束，此類倫理治理策略皆以要求系統設計者確保演算法的透明度和公平性。Howard 與 Borenstein（2018）提出的參與式設計方法，即是希望 AI 專業者自律，並透過公眾參與機制包容多元意見並接受共同監督，以避免偏見所造成的社會不平等的問題。「社會在環內」的 AI 也是基於讓公眾社會參與人機協作決策環的構想，作為達到包容多元、公開、透明、公平且負責任的 AI 典範（Rahwan, 2018）。

雖然各國組織已提出許多有關倫理準則規範的文件，仍可見多為宣示性質且鮮少有良好的實踐，因此有學者建議從設立標準作為管制策略的角度，提出概念性的 AI倫理治理五大支柱（Winfield & Jirotka, 2018），當作獲取公眾信任的溝通語言。Floridi（2018）則從倫理、管制與治理三者之間具有交互關係的前提下，建議應瞭解公眾對 AI 科技的接受度和使用度，提出前瞻分析週期環和倫理影響評估的模式。倫理影響評估亦為公民參與科技治理的可行方案，未來研究調查的重點也應該從公眾對 AI 科技的接受度和使用意願著手。

從文獻探討得到的啟發，建議我國政府可針對最受關注的 AI 倫理議題研議治理框架，確立「以人爲本」的中心思想，應對 AI 倫理、技術、與相關法律的挑戰，讓 AI 更安全並盡責地服務於人群、監測勞動市場的變化以及社會經濟的變革。增加公眾關注 AI 科技治理，並參與決策，將可鬆綁監限範圍，推廣可信賴的 AI，實現尊重人類尊嚴、具多元性和包容性的科技社會。藉助開發 AI 用於臺灣產經學研各界，或

可能有潛在倫理疑慮之項目，將之預設在 AI 倫理框架中訂定規範。在監管方面，有效實施與執行現有國家法規並釐清法規之限制，設計出的 AI 系統功能應具有可修正性，且有效分配不同權益關係者間之責任，如此才能掌握 AI 帶來的新興風險之因應對策。

最後以 Floridi 等人（2018）提出「良好 AI 社會」的倫理架構，所使用的四個標題重點作為本文的結尾：「我們可以成為誰：使人類自我實現，而不貶低人類的能力。我們能做什麼：加強人類的組織行為，而不消除人類的責任。我們可以實現的：在不減少人類控制的情況下提高社會能力。我們如何互動：培養社會凝聚力，同時不削弱人類的自決。」（pp.691-693）又如美國前總統林肯曾說：「預測未來的最好方式，就是去創造未來。」所以人類面對 AI 社會的未來，就端看我們自己如何努力了！

參考文獻

上村惠子、小里明男、至賀孝広、早川敬一郎（2018.6）。〈日米欧の地域特性に着目した AI 倫理ガイドラインの比較〉。「人工知能学会全国大会論文集 2018 年度人工知能学会全国大会 (第 32 回) 」 論文集。一般社団法人人工知能学会。

日本總務省情報通信政策研究所（2016）。《AIネットワーク化検討会議 報告書 2016 の公表—「AIネットワーク化の影響とリスク—智連社会（WINS(ウインズ)）の実現に向けた課題」》。取自 http://www.soumu.go.jp/menu_news/s-news/01iicp01_02000050.html

柯志明（2011）。〈應如何對待動物——對動物倫理之基礎與原則的一個反省〉。《應用倫理評論》，51：105-122。

劉育成（2020）。〈如何成為「人」：缺陷及其經驗作為對人工智能研究之啟發—以自動駕駛技術為例〉。《資訊社會研究》，38：93-126。

鍾張涵（2019.08.28）。〈當 AI 變同事〉。《天下雜誌》，680：72-77。

Adamson, G., Havens, J. C., & Chatila, R. (2019). Designing a value-driven future for ethical autonomous and intelligent systems. *Proceedings of the IEEE, 107*(3), 518-525. doi:10.1109/JPROC.2018.2884923

Adnan, N., Md Nordin, S., bin Bahruddin, M. A., & Ali, M. (2018). How trust can drive forward the user acceptance to the technology? In-vehicle technology for autonomous vehicle. *Transportation Research Part A: Policy and Practice, 118*, 819-836. doi:10.1016/j.tra.2018.10.019

Anderson, M., & Anderson, S. L. (2011). *Machine ethics*. New York: Cambridge University Press.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., ...Rahwan, I. (2018). The Moral Machine experiment. *Nature, 563*(7729), 59-64. doi:10.1038/s41586-018-0637-6

Balthazar, P., Harri, P., Prater, A., & Safdar, N. M. (2018). Protecting your patients' interests

in the era of big data, artificial intelligence, and predictive analytics. *Journal of the American College of Radiology, 15*(3), 580-586. doi:10.1016/j.jacr.2017.11.035

Boddington, P. (2017). *Toward the codes of ethics for artificial intelligence*. Switzerland: Springer International Publishing.

Borenstein, J., Herkert, J. R., & Miller, K. W. (2019). Self-driving cars and engineering ethics: the need for a system level analysis. *Science and Engineering Ethics, 25*(2), 383-398. doi: 10.1007/s11948-017-0006-0

Brutzman, D., Blais, C. L., Davis, D. T., & McGhee, R. B. (2018). Ethical mission definition and execution for maritime robots under human supervision. *IEEE Journal of Oceanic Engineering, 43*(2), 427-443. doi: 10.1109/JOE.2017.2782959

Bryson, J. J. (2018). Patiency is not a virtue: the design of intelligent systems and systems of ethics. *Ethics and Information Technology, 20*(1), 15-26. doi:10.1007/s10676-018-9448-6

Cath, C., Wachter, S., Mittelstadt, B., Taddeo, M., & Floridi, L. (2018). Artificial Intelligence and the "Good Society": the US, EU, and UK approach. *Science and Engineering Ethics, 24*(2), 505-528. doi:10.1007/s11948-017-9901-7

Chakraborty, S. (2018). Can humanoid robots be moral? *Ethics in Science and Environmental Politics, 18*, 49-60. doi:10.3354/ESEP00186

Chomanski, B. (2019). What's wrong with designing people to serve? *Ethical Theory and Moral Practice*, 1-23. doi: 10.1007/s10677-019-10029-3

de Swarte, T., Boufous, O., & Escalle, P. (2019). Artificial intelligence, ethics and human values: the cases of military drones and companion robots. *Artificial Life and Robotics*. doi:10.1007/s10015-019-00525-1

De Winter, G. (2018). Ai personalities: Clues from animal research. *Journal of Experimental and Theoretical Artificial Intelligence, 30*(5), 547-559. doi:10.1080/0952813X.2018.1430861

Di Nucci, E. (2019). Should we be afraid of medical AI? *Journal of Medical Ethics, 45*(8), 556-558.

Droste, W., Hoffmann, K. P., Olze, H., Kneist, W., Krüger, T., Rupp, R., & Ruta, M. (2018). Interactive implants: Ethical, legal and social implications. *Current Directions in Biomedical Engineering, 4*(1), 13-16. doi:10.1515/cdbme-2018-0004

Ema, A., Osawa, H., Saijo, R., Kubo, A., Otani, T., Hattori, H., ... Ichise, R. (2019). Clarifying privacy, property, and power: Case study on value conflict between communities. *Proceedings of the IEEE, 107*(3), 575-581. doi:10.1109/JPROC.2018.2837045

Etzioni, A., & Etzioni, O. (2016). AI assisted ethics. *Ethics and Information Technology, 18*, 149-156. doi: 10.1007/s10676-016-9400-6

Fiske, A., Henningsen, P., & Buyx, A. (2019). Your robot therapist will see you now: ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. *Journal of medical Internet research, 21*(5), e13216. doi: 10.1007/s10677-019-10029-3

Floridi, L. (2018). Soft ethics, the governance of the digital and the General Data Protection Regulation. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 376*(2133). doi:10.1098/rsta.2018.0081

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... Vayena, E. (2018). AI4People-An ethical framework for a good ai society: opportunities, risks, principles, and recommendations. *Minds and Machines, 28*(4), 689-707. doi:10.1007/s11023-018-9482-5

Future of Life Institute (2017). *Asilomar AI principles*. Retrieved from https://futureoflife.org/ai-principles/

Gill, A. S. (2019). Artificial intelligence and international security: the long view. *Ethics and International Affairs, 33*(2), 169-179. doi:10.1017/S0892679419000145

Green, B. P. (2018). Ethical reflections on artificial intelligence. *Scientia et Fides, 6*(2), 9-31. doi:10.12775/SetF.2018.015

Hassabis, D. (2017). Artificial intelligence: Chess match of the century. *Nature*, 544, 413-414. doi: 10.1038/544413a

Hoffmann, C. H., & Hahn, B. (2019). Decentered ethics in the machine era and guidance for AI regulation. *AI and Society, 35*, 635-644. doi: 10.1007/s11948-017-9975-2

Howard, A., & Borenstein, J. (2018). The ugly truth about ourselves and our robot creations: The problem of bias and social inequity. *Science and Engineering Ethics, 24*(5), 1521-1536. doi:10.1007/s11948-017-9975-2

IEEE (2016). *The IEEE global initiative on ethics of autonomous and intelligent systems*. Retrieved from https://standards.ieee.org/industry-connections/ec/autonomous-systems.html

IEEE Global Initiative (2019). *Ethically aligned design: a vision for prioritizing human well-being with autonomous and intelligent systems*. Retrieved from: https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf

Ienca, M. (2019). Democratizing cognitive technology: a proactive approach. *Ethics and Information Technology, 21*(4), 267-280. doi:10.1007/s10676-018-9453-9

Johnson, D. G., & Verdicchio, M. (2019). AI, agency and responsibility: the VW fraud case and beyond. *AI & Society, 34*(3), 639-647. doi: 10.1007/s00146-017-0781-9

Kamishima, Y., Gremmen, B., & Akizawa, H. (2018). Can merging a capability approach with effectual processes help us define a permissible action range for AI robotics entrepreneurship? *Philosophy of Management, 17*(1), 97-113. doi:10.1007/s40926-017-0059-9

Kanuck, S. (2019). Humor, ethics, and dignity: Being human in the age of artificial intelligence. *Ethics and International Affairs, 33*(1), 3-12. doi:10.1017/S0892679418000928

King, T. C., Aggarwal, N., Taddeo, M., & Floridi, L. (2019). Artificial intelligence crime: an interdisciplinary analysis of foreseeable threats and solutions. *Science and Engineering Ethics*. doi:10.1007/s11948-018-00081-0

Kirkpatrick, J. N., & Pearlman, A. S. (2019). Ethical challenges in the practice of echocardiography: What Is right and how do we do it? *Journal of the American Society of Echocardiography, 32*(2), 233-237. doi:10.1016/j.echo.2018.09.015

Kroll, J. A. (2018). The fallacy of inscrutability. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 376*(2133). doi:10.1098/rsta.2018.0084

Kurzweil, R. (2005). *The singularity is near: when humans transcend biology*. New York: Viking.

Lara, F., & Deckers, J. (2019). Artificial intelligence as a Socratic assistant for moral enhancement. *Neuroethics*. doi:10.1007/s12152-019-09401-y

Legg, S., & Hutter, M. (2007). A collection of definitions of intelligence. In B. Goertzel, & P. Wang (eds.), *Advances in artificial general intelligence: concepts, architectures and algorithms*. Amsterdam, Netherlands: IOS Press.

Levy, D. (2009). The Ethical treatment of artificially conscious robots. *International Journal Social Robot*, 1, 209-216.

Livingston, S., & Risse, M. (2019). The future impact of artificial intelligence on humans and human rights. *Ethics and International Affairs, 33*(2), 141-158. doi: 10.1017/S089267941900011X

Lu, H., Li, Y., Chen, M., Kim, H., & Serikawa, S. (2017). *Brain intelligence: Go beyond artificial intelligence*. Cornell University Library. Retrieved from https://arxiv.org/abs/1706.01040

Mendling, J., Decker, G., Reijers, H. A., Hull, R., & Weber, I. (2018). How do machine learning, robotic process automation, and blockchains affect the human factor in business process management? *Communications of the Association for Information Systems, 43*(1), 297-320. doi:10.17705/1CAIS.04319

McDougall, R. J. (2019). Computer knows best? the need for value-flexibility in medical AI. *Journal of Medical Ethics, 45*(3), 156-160. doi:10.1136/medethics-2018-105118

Miller, A. (2019). The intrinsically linked future for human and Artificial Intelligence interaction. *Journal of Big Data, 6*(1), 38. doi: 10.1186/s40537-019-0202-7

Rahwan, I. (2018). Society-in-the-loop: programming the algorithmic social contract. *Ethics and Information Technology, 20*(1), 5-14. doi:10.1007/s10676-017-9430-8

Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G., & Chin, M. H. (2018). Ensuring fairness in machine learning to advance health equity. *Annals of Internal Medicine, 169*(12), 866-872. doi:10.7326/m18-1990

Russell, S. J., & Norvig, P. (2003). *Artificial intelligence: a modern approach* (2nd ed.). Upper Saddle River, New Jersey: Prentice Hall.

Rybarczyk, Y., Cointe, C., Gonçalves, T., Minhoto, V., Deters, J. K., Villarreal, S., Gonzalo, A. A., Baldeón, J., & Esparza, D. (2018). On the use of natural user interfaces in physical rehabilitation: A web-based application for patients with hip prosthesis. *Journal of Science and Technology of the Arts, 10*(2), 15-24. doi: 10.7559/citarj.v10i1.402

Schönberger, D. (2019). Artificial intelligence in healthcare: a critical analysis of the legal and ethical implications. *International Journal of Law and Information Technology, 27*(2), 171-203. doi: 10.1093/ijlit/eaz004

Secretary of Defense for Acquisition Technology (1998). *DoD modeling and simulation (M&S) glossary*, DoD 5000.59-M. Retrieved from https://web.archive.org/web/20070710104756/http://www.dtic.mil/whs/directives/corres/pdf/500059m.pdf

Shank, D. B., DeSanti, A., & Maninger T. (2019). When are artificial intelligence versus human agents faulted for wrongdoing? Moral attributions after individual and joint decisions. *Information Communication and Society, 22*(5), 648-663. doi: 10.1080/1369118X.2019.1568515

Solovyeva, A., & Hynek, N. (2018). Going beyond the “killer robots” debate: Six dilemmas autonomous weapon systems raise. *Central European Journal of International and Security Studies, 12*(3), 166-209.

Turner Lee, N. (2018). Detecting racial bias in algorithms and machine learning. Journal of Information, *Communication and Ethics in Society, 16*(3), 252-260. doi:10.1108/JICES-06-2018-0056

Umbrello, S., Torres, P., & De Bellis, A. F. (2019). The future of war: could lethal autonomous weapons make conflict more ethical? *AI and Society*. doi:10.1007/s00146-

019-00879-x

UNESCO (2018). Artificial intelligence: The promise and the threats. *The UNESCO Courier*, July-September. Retrieved from: https://en.unesco.org/courier/2018-3

Wallach, W., & Allen C. (2009). *Moral machines: Teaching robots right from wrong*. New York: Oxford University Press.

Weber, A. S. (2018). Emerging medical ethical issues in healthcare and medical robotics. *International Journal of Mechanical Engineering and Robotics Research, 7*(6), 604-607. doi:10.18178/ijmerr.7.6.604-607

Winfield, A. F. T., & Jirotka, M. (2018). Ethical governance is essential to building trust in robotics and artificial intelligence systems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 376*(2133). doi:10.1098/rsta.2018.0085

Wirtz, J., Patterson, P. G., Kunz, W. H., Gruber, T., Lu, V. N., Paluch, S., & Martins, A. (2018). Brave new world: service robots in the frontline. *Journal of Service Management, 29*(5), 907-931. doi:10.1108/JOSM-04-2018-0119

Wright, S. A., & Schultz, A. E. (2018). The rising tide of artificial intelligence and business automation: Developing an ethical framework. *Business Horizons, 61*(6), 823-832. doi: 10.1016/j.bushor.2018.07.001

Zanzotto, F. M. (2019). Human-in-the-loop Artificial Intelligence. *Journal of Artificial Intelligence Research, 64*, 243-252. doi: 10.1613/jair.1.11345