doi: 10.29843/JCCIS.202307_(45).0002

為何應該以人工智能強化倫理衝突的緊急決策？

甘偵蓉

摘要

針對緊急且涉及倫理兩難或衝突而人們難以做出良好決策的事項上，本文主張應該設計人工智能（artificial intelligence，簡稱 AI）來改善甚至取代人們做倫理決策。這種 AI 倫理衝突決策系統，表面上是取代人類做出決策，但實質上是人類借助 AI 做出較佳的倫理決策，這種決策是從資料驅動到 AI 驅動。本文透過檢討那些反對以 AI 做倫理決策的方法論與本體論等兩類批評，並從檢討中指出這兩類批評，有哪部分值得 AI 開發者警惕，有哪部分則是對於這類倫理決策系統的誤解。本文並進一步指出，這種借助 AI 來改善涉及公共事務的緊急倫理衝突決策，若能從 AI 系統設計到部署的每個階段，皆納入決策利益相關人員甚至公眾的參與，便是辛諾特—阿姆斯壯與斯科堡於 2021 年所主張的「人工改良民主」。

⊙ 關鍵字：人工智能、倫理決策、緊急情況、倫理兩難／衝突、公共參與及民主

⊙ 本文作者甘偵蓉為國立清華大學人文社會 AI 應用與發展研究中心博士後研究學者。

⊙ 聯絡方式：Email ：gan.rrec@gmail.com；通訊處：300 新竹市東區光復路二段101號。

⊙ 收稿日期：2022/10/25　接受日期：2023/03/06

⊙ 本文感謝本刊兩位匿名評審的寶貴意見、丁川康教授、謝世民教授、許漢教授對於本文撰寫的啟發、曹家榮教授的鼓勵，也謝謝國科會計畫補助（計畫代碼：MOST 109-2423-H-007-002-MY2）。

Why AI Should Augment Urgent Decision-Making Involving Ethical Conflicts?

Zhen-Rong Gan

Abstract

This article argues that artificial intelligence (AI) should be designed to improve or even replace human decision-making in urgent situations involving ethical dilemmas or conflicts, which are well-known to be difficult for people to make good decisions. While the AI ethical conflict decision-making system may seem to supplant human decision-making, in reality, humans utilize AI to make better ethical decisions, transitioning from data-driven to AI-driven decision-making. By examining two types of criticisms—methodology and ontology—against using AI in ethical decision-making, this article points out which parts AI developers should be aware of and which are misunderstandings of these systems. Furthermore, the article suggests that involving decision-makers and even the public in every stage of AI system design and deployment can be viewed as an "Artificial Improved Democracy" proposed by Sinnott-Armstrong and Skorburg in 2021, which could enhance the development of AI ethical decision-making systems involving public affairs.

⊙ Keywords: Artificial Intelligence (AI), Ethical decision-making, Urgent situations, Ethical dilemmas／coflicts, Public participation and democracy

⊙ The author, Zhen-Rong Gan is Postdoctoral Research Scholar in Research center for AI development and Applications in humanities and social sciences at National Tsing Hua University

⊙ Corresponding author: Zhen-Rong Gan, email: gan.rrec@gmail.com; address: 101, Section 2, Kuang-Fu Road, Hsinchu 300044, Taiwan R.O.C.

⊙ Received: 2022/10/25　Accepted: 2023/03/06

⊙ I am grateful to the anonymous reviewers for their professional and insightful comments. My appreciation also extends to Professors Chuan-Kang Ting, Sher-Min Shei and Hahn Hsu, whose work inspired this paper. I would also like to thank Professor Chia-Rong Tsao for his encouragement. This work was supported by the National Science and Technology Council (Project code: MOST 109-2423-H-007-002-MY2).

壹、前言

去年台灣發生一輛公車撞死一名高中生事件，起因是該公車疑似煞車失靈，司機供稱當時雖看到前方有 10 多名高中生正依據綠燈指示穿越馬路，但「當下只有 2 個選擇，一是讓公車撞山壁，二是繼續往前開，因車上有 40 多名乘客，所以只能往前開」（中央社，2021）。

若我們不論這起事件調查結果司機所言是否屬實，任何人處在司機的情境中都很難抉擇，或許會和司機做一樣的判斷，但也許不一樣。諸如這種決策情境在哲學上稱為倫理兩難（ethical dilemma）：意即決策者當下只有兩種決策方式，而無論選擇哪一種，都必定有人會因為決策者沒選擇的那項決策方式而受到傷害；即使他們受到傷害的原因，絕非決策者蓄意使然，但與決策者所做的選擇有因果關聯，屬於該選擇下無可避免會產生的惡果，且該惡果與選擇下所意圖達到的善果相比，並不會不成比例。換言之，所謂倫理兩難決策，如果決策的行動本質是善的、不論選擇或放棄都是出於善的動機與手段、且所產生的善果與附帶的惡果不會不成比例等三條件皆具備的話，相關決策結果就會出現雙重效應（double effect）（McIntyre, 2019）。

其實倫理兩難情境本來就很難抉擇，若再加上情況危急，不容許有仔細思考的時間，多數人都只能憑直覺驟下決定，儘管人們不太會苛責決策者此時所做的決策，但在此情境下所做的決策不但難有品質，且未必符合決策者在深思熟慮下所認可的倫理原則或價值。

當然所謂有品質的決策，雖然有很多不同的評估指標，但基本上多屬於風險／影響評估（risk／impact assessment）的形式：亦即針對可達成預定目標下各項可行的決策，逐一列出每項決策所可能導致的後果，以及各利益相關人員在每項決策下可能受到的影響，然後從中選擇一個既能達成目標又能讓各利益相關人員受到最小負向影響的決策。

但在緊急的倫理衝突決策情況下，多數人很可能受到如時間限制、環境壓力、資訊限制等客觀因素影響，或是受到如疲倦、偏見、遺忘、緊張、疏忽、注意力不集中等主觀因素影響，以至於在當下所作出的決策，很有可能與去除前述主客觀因素的理想情況下所做的倫理決策不同。這裡的主客觀因素是指，那些人們對於自身做了

錯誤判斷常會歸咎的因素，或是人們認為若能去除掉將是比較理想決策情境的因素（Sinnott-Armstrong & Skorburg, 2021; Richardson, 2018）。

本文指出去除那些主客觀因素所做出來的倫理決策是較佳或較理想的，既不意味理想或較佳的倫理決策是不包含情緒、感知或情感，也不意味目前以機器學習所設計的 AI 因為缺乏情緒、感知或情感等以至於所做的決策較佳。因為一方面除了理性外，情緒與社會學習皆可能是人類做道德判斷的重要構成因素（De Sousa, 1979; Kauppinen, 2022）；[1] 另方面當不少人認為 AI 缺乏情感能做較佳決策時，通常心裡想的是缺少衝動或情緒化等這類一般認為會阻礙個人審慎思考的非理性因素，而不一定甚至不會排除同情、同理、正義感等這類正向情緒。

此外，去除上述那些主客觀因素，儘管一般認為所做出來的倫理決策是較佳或較理想的，但不保證人們針對緊急倫理衝突的事項就一定能做出良好決策。這是因為所需決策事項可能涉及相當多元且複雜的倫理考量，要在緊急情況下做出正確或至少適當的倫理判斷，已超出人類限制，以至於往往只能憑個人直覺來決策。但目前已有不少實驗證明，直覺是不可靠的倫理判斷依據，不僅容易受到各種認知偏見的影響，甚至連瑣碎或不相干事物亦可能影響當下的倫理判斷（Knobe & Nichols, 2017）。[2]

有鑑於此，針對那些涉及倫理衝突且緊急的決策事項，不論是人們受到前述那些主客觀因素影響而很難做出良好倫理決策，還是因為涉及的倫理考量太過複雜而多只能倚賴不可靠的直覺來做決策，本文主張人類應該要開發人工智能（artificial intelligence，簡稱 AI）倫理決策系統來提升（enhance）人類在這類事項上的決策品質。但開發目的，表面上是以 AI 取代人類在這類事項上的決策，實質上是透過事先蒐集人們認為在這種兩難決策事項上應該納入考量的因素／特徵有哪些，然後藉助 AI 演算法擅長找出那些因素／特徵的複雜關聯性以及生成最佳化的預測／決策之特點，以幫助人類改善在這類事項上的決策品質。

此處以 AI 幫助，主要是指直接執行 AI 所生成的倫理決策，但亦不排除是由 AI

1. 此點感謝審查人的提醒。
2. 此點感謝審查人的提問使作者有機會思考的更清楚。

篩選出幾項倫理決策後供人類挑選與執行。前項做法雖然最易引起爭議及憂慮，但這正是本文撰寫目的，主要就在於釐清及回應那些爭議及憂慮，相關論述可參見本文第貳～肆節。至於後項做法雖看似目前常見的 AI 使用方式，然本文所討論的緊急倫理衝突決策事項多涉及公眾事務或公共政策擬定，後項做法正是主張 AI 應該介入某些民主決策程序的可行方式，相關論述可參見本文第肆節的第二小節之（一）人工改良民主與（二）倫理兩難決策的公共參與基礎。

諸如本文所討論的這類緊急且涉及倫理衝突的 AI 決策系統，儘管尚在初探階段（甘偵蓉、許漢，2020），但事實上如自駕車或類似的倫理衝突／兩難決策，不論是從技術還是從概念層面的討論，都已累積到自成一特定領域的討論（Cervantes, López, Rodríguez, Cervantes, Cervantes & Ramos, 2020）。本文擬將這些涉及倫理衝突／兩難的批評區分為，屬於方法論還是本體論的批評，並加以檢討及釐清。若相關批評屬於方法論上的問題，則可朝向在研究方法上做出適當的倫理設計等來改善。但批評若屬於本體論上的問題，本文將指出很可能多出自於對 AI 倫理兩難決策系統的錯誤認識與定位。本文並進一步倡議，這類倫理決策系統從開發到部署的每個階段，為何應該盡可能納入利益相關人員甚至民眾的參與。

這麼一來，如果緊急且有倫理衝突的決策事項，或者應該直接執行 AI 所生成的倫理決策，或者應該使用 AI 所篩選的幾項倫理決策來讓公眾挑選與執行，這類倫理決策便可視為是從資料驅動到 AI 驅動的決策。所謂 AI 驅動，主要是強調那些決策來自或借助 AI 的計算，無需限定只能由 AI 而不能是人類來執行決策。[3]

3. 審查人指出 AI 丟出的結果，如果只是人類做決策的一個「資料」，如此是否還能說是 AI 驅動，不無疑問。但作者認為如文中所指出的，將相關決策稱為 AI 驅動，主要是強調那些決策出自或有借助 AI 的計算（Colson, 2019），在理解上或可比擬那些「以實證為基礎的決策」（evidence-based decisions），簡言之，若將 AI 驅動改稱為以 AI 為基礎的決策也可以。另外審查人舉例指出，人類很多行為和決定受到社群媒體演算法的影響，但我們似乎不太會說相關決定是「演算法驅動」的？然而作者認為，如果我們可接受人類的某些行為和決策是受到「演算法的操弄」這種說法，如將「」的詞置換成「演算法的驅動」似無不可。

貳、人工智能倫理決策系統

所謂需要做出倫理兩難／衝突的決策情況，儘管現實中罕見上述公車司機所遇到的情況，但針對不一定涉及生死抉擇的其他兩難道路駕駛情境，其實在日常生活中隨處可見。例如，在高速公路上因為路況不熟或一時疏忽，而幾乎要錯過交流道的情況下，此時得迅速觀察與決定，是否繼續往前開而可能因此錯過參與一個相當重要的會議，或是緊急變換車道下交流道去，而可能增加後方車輛來不及減速而追撞上來的危險。又或者上班時間快來不及時，開車至人車眾多的十字路口，正巧遇到閃黃燈，此時得迅速觀察與決定，究竟要踩油門加速通過，而可能增加不慎撞上另一幹道的車輛或行人搶綠燈通行的風險，或是踩剎車停下但自己上班遲到，且可能增加後方車速頗快的車輛煞車不及追撞上來的危險。

諸如這些情境，駕駛人在不違反交通規則的情況皆有判斷與選擇的空間，而究竟是否判斷正確且採取不危及自身與他人安全的行動，既考驗駕駛人的觀察與判斷能力，也考驗駕駛人的反應速度及行車技術。當然也考驗其他用路人的反應與行車技術，例如，行人注意到有輛車快速駛來，可能會確定該車有減速且預備停下的跡象，才會準備過馬路，而不會僅憑號誌轉換就冒然通行；或是後方來車駕駛人看到前方車輛疑似準備變換車道或減速，可能會稍加減速以因應前方車輛在行車方向或速度上的改變。換言之，道路交通安全基本上是所有用路人協作下才有可能達成的目標，任何一方的疏失都有可能發生交通事故。

暫且擱置自駕車與人駕車能否順利協作的問題（劉育成，2020），[4] 諸如上述需要駕駛人緊急判斷與採取行動的倫理兩難或衝突情境，若我們能將在這類情境下可能

4. 劉育成（2020）指出，AI 技術儘管形塑了人類智能與 AI 如何共存，但只要 AI 持續以窮盡外在環境資訊與情況為邏輯的話，不但 AI 要像人類一樣思考還有相當長一段路得走，且更有可能是以犧牲人類智能為代價才有可能達到。例如自駕車無法準確執行，歸咎原因通常指向行人不遵守號誌、不走斑馬線、不依照規定來行動等。另該文從技術人類學與現象學的角度探討 AI 與人類智能的異同，以及 AI、人類、環境三者交互作用的關係。分析細緻且相當具啟發性，但與本文純粹將 AI 視為協助人類找出倫理衝突／兩難情境下的最佳決策工具，這種工具應該被如何看待及生產過程中應該如何規範，才有可能生產出適當的倫理決策等，在探討方向上頗不同於劉文，故不深入討論。

應該採取的倫理決策模式全找出來，事先內建在自駕車或人駕車的駕駛輔助系統當中，一旦遇到倫理兩難或衝突情境，自駕車就能依據內建模型來做決策，或是即時提供駕駛人決策建議。這麼一來，相較於目前倚賴個別駕駛人當下的判斷、反應以及駕駛能力，是否更有效降低交通事故發生的機率或事故傷亡人數，更有助於促進交通安全這項終極目標？

事實上不只道路駕駛，舉凡決策事項涉及必須在人類生命、財產、安全或自由、動物生命與福祉、環境安全與永續、社會共善與福祉等人們所重視的各種利益（interests）之間緊急做出選擇與比較，但受到時間、體能、思維模式、資訊獲得不全、繁雜或過多等因素影響，以至於人類通常很難在這種情況下做出良好的倫理決策。換言之，人類針對某些相當重要但必須緊急做倫理決策的事項，基於前述主客觀因素，人們通常難以做出良好決策，儘管相關決策結果未必就是錯的或惡的，也有可能還是做出自己所能接受的決策結果。然一旦人們拒絕自己最後所採取的決策，往往就是歸咎前述那些主客觀因素的存在所導致的疏失或過失等。

再者，何謂良好的倫理兩難決策，因為涉及價值或倫理原則的衝突，並沒有先驗或顯而易見是正確的決策選項可供挑選，還可能涉及決策者與受決策影響者等人員、文化、時地等情境脈絡的差異，而造成決策結果上的差異。

在這情況下，以人類經驗為師，從過往大量相關經驗當中找出並學習較佳的倫理決策選項，這正是當代主流設計人工智能的機器學習特性，或可用來嘗試解決倫理兩難決策在現實中決策品質通常不佳的問題。畢竟堅持只倚賴人類當下自身的倫理判斷且無論好壞都應該接受的主張，卻拒絕利用科技幫助人類所重視的那些利益獲得更多保護，究竟有多少說服力頗令人懷疑。

所以，人類借助 AI 找出在緊急情況下通常不擅長做出良好倫理決策的事項上之最佳決策，尤其利用機器學習能自動且迅速處理巨量資料的特點，以改善人類在特定事項上需緊急採取決策或行動的倫理判斷品質，這件事值得受到更多重視與努力。在此同時，無需也不應該宣稱 AI 已發展出如人類一樣有做各種倫理決策的道德能力，更不必陷入 AI 可不可能或應不應該做倫理決策的論辯。未來不論是否真有可能研發出能力與人一樣的通用 AI（artificial general intelligence）或強 AI（strong artificial intelligence），而值得探討這類 AI 能否如人一樣發展出有能力做各種倫理決策的道

德主體，目前主張以特定功能的弱 AI（artificial narrow intelligence or weak artificial intelligence）來提供有關特定事項的最佳倫理決策，可以純粹指設計一個在功能上可處理及產出特定決策資訊的 AI，而無任何本體或形上學的承諾這類 AI 足以被視為有倫理決策的能力。[5]

釐清為何要以 AI 來處理特定事項的倫理兩難決策，以及現階段應該如何看待研發這類 AI 系統的目的，不但是回應近年來當人們發現日常生活有越來越多地方使用 AI 時，人們對於 AI 的看法常出現兩極化的現象：或者過度期待 AI 能做到目前還是只有人類才能做到的事，或者過度擔憂 AI 已經取代原本以為只有人類才能做到的事。上述釐清，也有助於我們從 AI 有無能力或是否應該做倫理決策的論辯泥沼中脫困，改將焦點放在有關提供特定事項的 AI 倫理決策系統，究竟應該以哪些因素或特徵作為參考指標以及它們相互作用結果等方面的探討。

就如同臨床上被用來監測臥床病人在哪些情況下需要醫生緊急救治的心電監護診斷儀，在設計這種儀器時，就得決定要參考及監測哪些生理參數，並得決定當哪些參數超過預先設定的正常值範圍時，需要發出警示好讓醫護人員趕緊過來救治。若從擬人化角度來看心電監護診斷儀所展現的功能，也可以說該機器必須依據所輸入的數據特徵，在極短時間內迅速判斷並做出是否通知醫生前來救治的倫理決策。但基本上不會有人問這種機器有無能力或是否應該做這麼重大的倫理決策，因為該機器只是依照預先設定的內建程式運作，並無稱得上在做倫理決策時所應具備如自主與自由選擇的能力等。再者，若希望這種機器能更精準達成在最恰當時機警示醫護人員前救治的目標，更要緊的是檢視所設定監測的生理參數項目是否正確或足夠，以及相關警示閥值的設定是否需要調整等。

AI 系統雖然不若心電監護診斷儀，它有自主決策的空間，但該自主意涵與人類的自主意涵包含了自我設定目標且有自由選擇空間等，尚有非常遙遠的差距。尤其以

5. 有關不論是弱 AI 目前還是未來通用／強 AI 可能帶來的倫理問題，相關國外文獻頗多，若希望對此有一初步的整體掌握，推薦閱讀劉湘瑤、張震興、張瓅勻、趙恩、李思賢（2021）一文，該文整理了近期文獻共 82 篇，內容涵蓋了特定倫理議題及倫理政策治理，相當全面。

機器學習設計的 AI 系統，是指系統依據演算法所設定的搜尋方式，可自動比對目前所輸入的資料特徵與過往所學習到的資料特徵，是否有相似之處以及相似程度，然後輸出有對應到或最相似過往所學習到的資料特徵類別或數值。所以 AI 研發人員要給機器學習什麼資料特徵、那些資料特徵對應到什麼類別、或對應的數值所代表的意義等，便決定了所建立的 AI 系統能否達到預期目標或展現預期功能。

如以本文所談論能對於特定事項提供最佳倫理決策的 AI 系統來說，這類 AI 系統在實際應用時，將輸出哪些類別的倫理決策，便取決於當初所學習相關倫理決策的訓練資料，究竟如何定義決策所需考量的因素或資料特徵，以及如何定義各特徵在不同權重與順序組合下所對應的決策類別。所以探討與檢視輸入這類決策系統的訓練資料究竟定義了哪些資料特徵，以及所輸出的結果是否符合人們認為在深思熟慮下所會做出的決策，如何借助科技力量來提升人們在某項事項上的決策品質，現階段來說或許會比討論是否應該讓 AI 做倫理決策來得重要。

尤其這類 AI 所做的倫理決策既然屬於倫理兩難情境，這代表每項決策皆涉及對於不同重要價值或利益衝突的權衡（trade-off），而權衡結果也就是系統所輸出的結果，究竟對哪些群體有利或不利，尤其所不利的群體是否包含以下現象：多集中在歷史上曾構成歧視的如性別、膚色、外貌、身形、種族、社會階級等特定社會身份，但那些身份一般認為與這類倫理決策無關，以至於雖然未被列入當初輸入的訓練資料特徵中，但決策結果卻明顯不利於擁有那些身份的群體。當系統開發團隊發現有此現象，便需要回頭檢視系統當初所輸入的資料、演算法的設計與模型建立等，究竟哪個環節出了問題，以致於系統輸出結果疑似讓特定群體受到不平等對待。

不意外的是，利用機器學習來改善倫理兩難或衝突決策相關研究，多屬於初步嘗試階段，難免有錯誤而招致批評。但批評嘗試的研究方法錯誤，有需要改善是一回事；但批評這類決策不應該以 AI 來模擬，否則會讓本來就不易產出良好決策的特定事項變得更糟，則是另一回事。而辨識相關批評是在哪個層面上談論，並釐清這些批評有哪些值得留意，哪些可能有所誤解或不公允？如此一來或有助於推動人們思考，當試圖利用 AI 來提升某些事項涉及重大倫理兩難的決策品質，究竟是值得繼續努力的方向，還是根本搞錯了方向？

因此，以下二節擬將有關 AI 倫理兩難決策系統相關研究的批評，區分為方法論

以及本體論的批評。所謂方法論的批評是指，這類批評未必反對以 AI 來提昇某些事項的倫理決策品質，但反對資訊科學界目前常以眾包（crowdsourcing）方式來獲得輸入機器的訓練資料，並批評研究團隊在資料特徵的選擇上有複製社會偏見甚至歷史歧視之嫌。本體論的批評則是指，這類批評基本上反對以 AI 來提升某些事項的倫理決策品質，他們既不認為以量化計算為本質的機器學習有可能成功模擬倫理決策，也不認為透過蒐集群眾意見有可能獲得在倫理兩難情境下應該如何決策的答案。

參、方法論的批評與釐清

本節討論有關倫理兩難決策 AI 研究在方法論上的批評，又可區分為輸入資料眾包與資料特徵選擇不當，以下將分成兩小節來討論。第一小節將在簡介有關眾包的批評內容後指出，眾包是一種蒐集經驗資料的方法，如同過往社會科學中許多蒐集經驗資料的研究方法，皆有優缺點，只要研究分析時能切合所蒐集的資料特性及限制，以眾包資料驅動機器學習如何做倫理決策未必不行。第二小節則在簡介有關資料特徵選擇不當的批評內容後指出，究竟要選擇哪些資料特徵來驅動機器學習以符合決策系統開發目標，不只涉及技術問題，還涉及價值選擇與承諾，換言之，當研究團隊在取捨要納入或排除哪些資料特徵時，就是一種價值選擇，本文將同時建議應該以決策系統所提供的結果如何促進平等價值來引導，也就是以平等主流化觀點來擬定學習模型的設計、檢測與調整等策略。

一、不當使用眾包資料

以群眾外包資料來探討未來以 AI 作倫理兩難決策的研究，最有名的莫過於 2016 年美國麻省理工學院所設計的道德機器（moral machine）的眾包網站。[6] 它以哲學上著名的電車難題這個思想實驗作為設計原型，以一輛自駕車遇到煞車失靈而必須在直行與轉向其他車道之間做選擇，但選擇代價是以犧牲兩組行人或行人與乘客其中之一的

6. 請參考 MIT Moral Machine website，https://www.moralmachine.net/。

生命。該網站提供了十種語言讓無任何身份限制的網路使用者線上填答，以便蒐集他們在包含了十幾組不同選擇條件的這種兩難情境下的決策結果，並宣稱此研究目的在於讓社會、道路監管政府單位、車商等眾人了解，未來上路的自駕車如果要內建這類倫理兩難決策，來自至少 233 個國家或地區在一年半內就累積近四千萬人次的看法。該眾包網站推出後獲得不少關注，不但有媒體報導及部落格討論，在《自然》雜誌上一篇由該團隊介紹此項研究成果的文章，至今已獲得三百多篇學術論文引用（Awad, Dsouza, Kim, Schulz, Henrich, Shariff, Bonnefon, Rahwan, 2018）。

爾後我國清華大學研究團隊也自行設計了以繁體中文為主有關人工智能的倫理兩難決策眾包網站，在他們設計的四種倫理兩難決策的類型當中，有一種就是自駕車，該網站同樣也在短短一年就迅速收到三萬多筆資料。[7]

兩研究團隊的背景及研究目的並不相同：MIT 團隊的研究背景主要是由社會科學家所組成，研究目的在於呈現社會大眾期待自動駕駛車輛預先植入哪些倫理決策的看法，無意宣稱應該以眾包資料來設計自駕車的倫理決策系統；清大團隊的研究背景則是演算法設計專家，研究目的在於企圖設計能模擬常民道德（folk morality）的演算法，而透過眾包所蒐集的資料主要是作為測試案例，並非真要建立適合內建在自駕車上的倫理決策模型。但兩研究團隊尤其是 MIT，因為在研究主題上都涉及以「AI 做倫理決策」這項頗為敏感主題，而受到不少批評。

之所以敏感，不只因為碰觸到許多人的不安全感，像是 AI 是否擁有向來被視為人類專屬的倫理決策能力，更在於某些人的生命，竟以自駕車這類機器預先決定未來危急時刻是可被犧牲的，這似乎在預謀殺人。尤有甚者，建立自駕車的倫理兩難決策模型所需訓練資料，如果類似這兩團隊都來自眾包，在已知這類資料常複製歷史偏見的情況下，未來以這類資料作為機器學習的輸入資料，相關輸出結果將對於原已受到社會歧視的非主流或弱勢群體更加不利。

上述三項引發人們不安全感的來源，前二項在第一節已說明，人們的不安全感與擔憂在目前並無必要，且人類若能事先審慎評估遇到緊急狀況應該如何決策的選項內

7. 請參考人工智慧倫理學網站，https://aiethics.ml/index.php。

建在自駕車，總比倚賴個人當下憑直覺做選擇佳。當然這不表示應該在自駕車中內建屆時哪些人將被犧牲的決策模型，這樣在道德上確實有問題，有關這點將在第二小節有關資料特徵選擇的討論中釐清。

有關引發不安全感的最後一項，也就是眾包資料不應該作為 AI 倫理兩難決策系統的訓練資料，確實需要注意與釐清。因為前述已指出，現階段只具有特定功能的 AI，距離人類具有在各種不同事項上做倫理決策的能力還相當遙遠，而倫理兩難決策如果確實屬於有雙重效應的決策，相關決策所產生的惡果基本上就不應該視為蓄意殺害。

然而必須留意的是，雙重效應決策當初提出的倫理兩難情境，在設想上屬於偶需做出單次決策的情境，未必能擴充解釋也適用在每次類似決策都對於相同群體不利的重複決策情境。這表示相關 AI 開發團隊在系統正式部署前，若已得知系統輸出結果總是對特定群體不利，但未進一步檢視相關不利是否在道德上可被合理說明，例如在自駕車案例中，不利的如果總是未遵守交通規則的群體，便有可能獲得初步（prima facie）說明，但不利的如果總是高齡群體或男性群體，可能就難以獲得初步說明，而有透過自駕車蓄意殺害特定群體之嫌。

不過眾包資料就如同以往各種不同研究方法所蒐集的經驗資料，相關經驗資料特性總是受到蒐集方法的影響。像眾包資料如果是透過網路平台匿名蒐集，雖然蒐集速度快且數量多，但相對地也容易蒐集到不需承擔任何選擇後果而不加思索就輕率回答的資料，以至於所蒐集到的資料未必能真實反應填答者的想法，更別提與填答者在真實世界的作為一致。當然來自線上匿名填答的眾包資料，這麼多年來學術界也發展出減緩前述缺失的研究設計策略與統計方法，但體認這類資料的本質限制並適當解釋與應用，是使用這類資料來設計系統並部署的 AI 團隊基於科學整全性應該遵守的。

若依據上述，MIT 與清大兩團隊利用眾包資料的目的，都不在於說明或找出自駕車的倫理兩難決策系統「應該」如何設計。像 MIT 團隊是有意透過眾包方式，以找出不同地理區域的網路使用者有關自駕車作倫理決策的看法。即使他們所收到的資料，是否真的與自駕車有關，或純粹與道路駕駛的倫理兩難決策有關，頗令人懷疑。例如，若將他們網站上的題目及圖片改成「有一輛人類駕駛的車子因為煞車故障，您認為駕駛當下應該如何決策？」，或許有不少人所填寫的答案與目前在自駕車題目中所

填寫的答案一樣，也說不定。

至於清大團隊蒐集眾包資料的目的，如果正是為了設計能成功模擬常民道德的演算法，那麼所收到的眾包資料是否針對自駕車則無妨，但得留心填答者在倫理兩難與非兩難情境中，或線上填答與現實場景中，所考量的倫理特徵或原則是否一致與穩定。若不一致，屆時以演算法所模擬的是屬於哪一部份的常民道德，則必須釐清。例如，據此資料所模擬的毋寧是數位常民也就是網友的道德觀，且偏向直覺反應而非深思熟慮下的道德選擇；而這類道德觀究竟與一般人在現實生活中差距有多大，或是與有受過訓練的倫理學家的道德觀差距有多大，尤其網友常被認為帶有偏見與歧視是否果然如此等，諸如這些問題正好可藉由清大透過眾包所蒐集到的資料來回答與比較。

其實兩研究團隊的研究目的，如果都是想探討多數人在涉及自駕車的倫理兩難情境下是如何決策，那麼在確認填答者於兩難與非兩難情境下所做的選擇是否一致的研究設計，或許能以增加問答題項以及兩種情境交錯詢問等方式來確認。又線上填答與置身現實場景的差異，則目前已有利用擴增虛擬實境技術來讓個人仿佛置身在自駕車的決策情境等來彌補。

不過對於反對以眾包資料來探討自駕車倫理決策的批評者來說，他們或許認為其批評重點在應然面而非實然面，也就是不可能透過眾包找出自駕車「應該」如何做倫理決策的資料。若是如此，這項批評涉及的就不只是相關研究方法論而是本體論上的批評了。因為目前主流設計 AI 系統的機器學習，正是透過讓機器學習大量的人類經驗資料，以便能對新輸入的資料進行正確預測或分類，但正確的倫理決策如果不可能從經驗資料中獲得，這表示以機器學習來找出倫理衝突／兩難情境下的最佳倫理決策是錯誤的。有關這項批評將在第肆節再仔細處理。

二、不當選擇資料特徵

然而，正確的倫理決策就算有可能從經驗中學習，該如何確認過去有關特定事項的倫理兩難決策經驗中，哪些是道德相關因素而需要被納入考量？

上述舉例的 MIT 團隊所設計的「道德機器實驗」眾包網站，在這部分尤其受到許多批評。因為實驗是從以下七項特徵當中隨機分配約 1~3 個特徵在各個決策兩難

的場景中，提供填答者權衡與選擇：物種（人／貓狗）、年齡（年輕／老人）、性別（男／女）、體型（胖／瘦）、社經地位（運動員／行政人員／醫生／流浪漢）、拯救或犧牲的生命數量（在繼續直行車道或改變車道上的行人與行人或乘客與行人之比較）、行人有無遵守交通號誌等。在這七項特徵當中，有些特徵會是現實中道路駕駛遇到緊急狀況時，通常不被認為是道德相關因素而應該列入考量的。例如將性別與體型納入考量，難道是認為男性或體型較大者比女性或體型較小者耐撞嗎？不僅如此，這些特徵還強化了社會常見的刻板印象與歧視（Jaques, 2019; Bigman & Gray, 2020），例如社經地位高低被列為考慮特徵，意味著人不但可從外觀來判斷個人的職業或社會階級，還可用來判斷生命價值的高低。

甚至，研究團隊將這七項特徵設計成兩難決策的二擇一選項，似乎傳遞出有哪些特徵可以拿來權衡生命價值高低的訊息，而違背了多年來許多社會在人權與動物權方面的努力與平權倡議。例如，不論是基於國際人權宣言與許多國家相關人權法規，還是基於德國倫理委員會 2017 年所公布的〈自動駕駛與車聯網指引〉（Ethics Commission on Automated and Connected Driving），諸如性別、年齡、外表、社會階級等，皆被禁止作為區辨生命價值高低以至於可以被優先拯救或犧牲的決定因素，在決策上是屬於任意且道德無關因素。同樣地，人類福祉將永遠置於動物生命之上的價值觀，也是動物保護運動人士多年來反對且不斷教育民眾應該改變的。所以像 MIT 這類的 AI 相關實驗設計，似乎復辟了社會多年來努力在道德上取得進展過程中想要揚棄的價值觀。儘管 MIT 團隊宣稱他們從眾包網站收到的資料，目的在於呈現一般民眾對自駕車做倫理決策的看法，並不是要直接輸入自駕車。但不論是否要實際設計 AI 系統，諸如這類 AI 研究在某種意義上可視為企圖研發或有關「殭屍 AI」（zombie AI）的研究——指將過往被揚棄的價值觀不當內建在 AI 系統裡（Vallor, 2021）。

針對上述有關研究條件設計不當的批評，如放在以機器學習研發 AI 系統的相關研究脈絡中，其實與機器學習的資料特徵選擇有關。MIT 實驗（或許清大實驗也是）所受到的批評，正可凸顯以下二件事的重要性：一是輸入 AI 的資料在特徵選擇上，應該如何與 AI 系統的開發目標與期待功能一致，不只是技術問題，還有倫理目標的選擇與承諾（陳瑞麟，2020）；另一是系統開發團隊決定輸入 AI 哪些資料特徵，本身就是一種價值判斷與選擇，並非價值中立，而判斷與選擇的理由對於如何看待模型

的分類或預測結果，其實影響很大。

（一）特徵選擇受到技術與 AI 開發目標的影響

首先，當有研究團隊宣稱探討以 AI 解決倫理兩難的決策問題時，即便所解決的問題都屬於倫理兩難或衝突，但針對不同事項與目標所需設定的倫理考量因素，也就是所謂的資料特徵，就會不同。例如，有關自駕車倫理兩難決策系統的資料特徵選擇，將與器官移植分配決策系統的資料特徵選擇不同。但即使都有關自駕車的倫理兩難決策系統，在以交通安全作為所要達到的最高決策目標下，針對突然煞車失靈的自駕車，其決策問題訂為「要拯救／犧牲乘客還是行人」，就可能與「避免／減少任何人傷亡」，在資料特徵的選擇與計算上便有所不同。

兩者僅管都應該將乘客與行人兩組人員的人數列為資料特徵，但根據 MIT 的設計情境，工程師在該情境中所要解決的問題顯然是「要拯救／犧牲乘客還是行人」。但如果將決策問題修正為「避免／減少任何人傷亡」，那麼不僅是兩組人員的人數會被列入資料特徵，車子前方沒有出現人類或出現最少人類的道路位置，以及車輛本身可承受的撞擊力道等，都會是需要列入計算的資料特徵，以便在綜合計算車內外人數與環境等資料特徵後，最終所找出的決策會是在行人與乘客都不會或受到最少／輕傷亡的最高數值。

這樣或可理解，為何有不少程式工程師對自駕車倫理兩難問題頗爲反感（Kalra & Groves, 2017; Iagnemma, 2018; Furey & Hill, 2021; Dixon, 2020; van Wynsberghe & Robbins, 2019）。因為對他們而言，在設定自駕車的決策選項時，即使有些決策後果無可避免帶來人員傷亡，但堅持任何情境下的決策選項，都絕不包含以撞擊、犧牲、或殺害人類作為目標，也就是不論個人還是群體的生命，都不應該成為權衡的項目。即便承認減低人員傷亡數量在道路交通安全中常是相關的道德考量，但以一人還是五人傷亡作為權衡目標來設定車子應該直走還是轉彎，或是以直走還是轉彎所產生的結果可能減低多少人傷亡作為權衡目標，兩者是有差異的。目前許多 AI 倫理指引裡都以《世界人權宣言》或國際人權公約等作為綱領（Fjeld et al., 2020），並強調以人為本、公平對待與不歧視等基本價值或倫理原則的促進。儘管如何做到保護人權與具體實現那些基本價值或倫理原則，方法相當多元，但「絕不以人類作為目標」（no human targets），應該是 AI 相關開發人員必須遵守的最基本行為規範（ethics of

design），且是任何 AI 系統稱得上是倫理的設計（ethics in design）都應該滿足的條件。

（二）特徵選擇與模型建立的平等主流化

其次，輸入 AI 的資料在特徵選擇上，不只如上述受到 AI 決策目標所隱含的基本價值或原則所引導，研究團隊對於社會上常見的偏誤，像是性別、種族、階級、障礙等差異所導致的不平等對待，是否有敏感度，也會影響研究團隊決定要納入或排除哪些資料特徵（王道維，2021），以及如何解讀由那些特徵及其關係所建構的模型其輸出結果是否要修正或調整。換言之，機器學習從資料特徵選擇到模型產出結果，這整個資料分析、演算法設計、模型建構以及結果產出的過程，不單只是技術與科學整全性的考量，還有研究團隊的價值選擇與判斷。

針對相關模型所產出結果如何避免可能導致不平等，目前雖已發展不少偵測的技術，但本文建議針對輸入 AI 的資料特徵選擇宜採取平等主流化（equality mainstreaming）的策略與思維。就如同為促進性別平等而採取性別主流化之後，不但可讓相關公共政策在計畫與施行時，就會將不同性別所受到的影響列入評估分析，目前幾乎所有科學量化研究在蒐集樣本時，性別通常都被視為基本資料欄位來蒐集。當輸入 AI 的資料特徵採取平等主流化（equality mainstreaming）策略時，相關開發團隊就必須思考與檢視，是否有輸入系統的資料原本應該將性別、種族、階級、障礙等列為資料特徵，卻被忽略了？或者檢測模型產出結果，有無原本不應該但實際上卻在前述項目上有明顯差異的現象，而有調整模型的必要？

例如，AI 人臉辨識系統過去曾被發現，其應用結果可能讓原本已受到社會歧視的有色女性落入更不利處境。究其原因是這類系統的預測模型，在建立時所輸入的訓練資料嚴重缺乏有色族群尤其是女性的照片，以至於對女性有色族群的正確辨識率很低。所以當這類系統被應用到住宅或辦公大樓門禁解鎖、犯罪嫌疑者與警政資料庫的人臉比對等事務時，可能老被鎖在大樓外面不得其門而入，或被誤當嫌疑犯扣押而冤枉無辜，最終導致女性有色族群受到不平等對待的惡果（Gebru, 2020）。

此外，研究團隊對於造成社會歧視的那些差異與不平等若缺乏敏感度，也將難以察覺有些資料特徵雖然不是受歧視的那些特徵，卻可能是那些特徵的代理因素（proxies），以至於所建構的模型其決策結果，仍舊讓原本已受到歧視的那些群體持

續處於不利處境。例如，個人的性格描述、生活興趣、就讀學校等特徵，在過去社會長期存在性別偏見下，很有可能就成為性別的代理因素。或者心理健康、犯罪紀錄等特徵，在社會過去長期存在種族偏見之下，有色群體常落入低收入、低成就感、低自尊、警務過度稽查與監控等不利處境當中，而往往成為種族的代理因素（Braun, Broestl, Chou & Vandersluis, 2021）。

針對那些歧視案例的檢討，目前多認為 AI 開發團隊有性別盲、種族盲、階級盲之嫌，以至於未能意識到如果是開發有可能被廣泛應用的人臉辨識系統，那麼在建立模型的訓練資料上，就應該涵蓋越多膚色、五官組合、族群、性別等差異，屆時模型的容錯率／可包容性才會越佳。所以相關解決辦法大致如下。像是盡可能讓訓練資料的內容多元化，但這會受限於資料量是否足夠的問題；倡議建立模型卡（model card），在卡片上清楚標註輸入模型的訓練資料於地理區域、種族、性別等方面的資訊（Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, ID., Gebru, T., 2019），就如同食品標示一樣，以便模型使用者可了解模型的特性、限制條件與使用脈絡，降低模型被應用在錯誤環境而帶來未預期的負向結果。

不過若有些事項既存數據存在嚴重偏差或稀少等缺陷的話，那麼建立模型卡的做法，就對於設計出能做出良好建議的決策模型沒什麼幫助。例如美國黑人家庭相較於白人家庭，因為較少機會接觸器官移植相關資訊，即使根據統計黑人發生腎衰竭的比例高於白人 2~4 倍，但目前美國的器官移植分配資料庫紀錄，不論是器官捐贈者還是等待被捐贈者，黑人數量都遠少於白人，在已知種族是降低器官移植後身體排斥機率所應考量的因素，以至於黑人接受腎臟移植的機率將遠比白人小很多。因此，若以目前資料庫數據來開發器官移植分配 AI 決策系統，可想而知就會複製甚至加劇美國社會在器官移植分配這件事情上目前的種族不平等現象。

如果能以平等主流化觀點來看上述這類 AI 決策系統的開發與部署，在選擇資料特徵時，就會意識到種族這項特徵很可能應該列為資料特徵，以避免種族盲的效應，且思考是否需調整某些與醫療及社會階級有關的健康因素其排序和權重，以檢測所輸出的模型結果是否有助於改變黑人在相關事項上的不利處境，或至少未複製或加劇目前不平等的現況。又或者開發團隊可將排除種族為資料特徵所產出的模型結果，拿來比較種族是資料特徵之一的模型結果，若有差異且無法被合理說明，那麼就是這類決

策系統應該調整模型的方向。當然研擬如何提供黑人家庭更多器官捐贈相關醫療資訊，以逐漸增加黑人在這類決策系統的捐贈與受捐贈的資料量，將是改變黑人在器官移植分配這件事情受到不平等對待的根本之道。

如果目前的 AI 決策系統多以資料驅動的機器學習來設計，相關開發團隊究竟從資料當中提取哪些特徵來建立模型，有無符合平等主流化便很重要。以目標是建立「避免／減少任何人傷亡」的自駕車兩難或衝突決策系統來說，諸如性別、種族、階級、障礙就不應該成為系統決策所應考量的特徵，但以前述各項特徵尤其障礙這項特徵來檢測是否會明顯影響決策結果，將有助於檢視有無其他資料特徵成為阻礙平等的代理特徵。這麼做或許可視為繼 M. Mitchell 等人（2019）倡議模型卡來改善模型應用可能造成不平等的方案後，另一種同樣採取非技術路徑來推動 AI 促進社會平等的倡議。

肆、本體論的批評與釐清

針對研發 AI 倫理決策系統在本體論上的批評，主要有二：一是批評倫理決策的必要條件之一是倫理敏感度，但這種敏感度並非以量化計算的機器學習所能成功模擬與複製的（Véliz, 2021）；另一是同樣針對眾包資料的批評，但不若上述方法論的批評是抱怨這類資料常包含社會大眾很容易存在的偏誤缺失，而是根本否定正確的倫理決策可以透過調查大眾經驗與看法中獲得（Braun, Broestl, Chou & Vandersluis, 2021；祖旭華，2020）。

一、計算系統無法模擬倫理敏感度

有關以量化計算為本質的機器不可能模擬倫理敏感度的批評，一方面得釐清何謂倫理敏感度，另方面得釐清機器學習要模擬的是什麼倫理決策。

首先所謂的倫理敏感度，在後設倫理學的討論當中通常是指，決策者能認知或感受到決策情境中的道德相關因素，並且能隨著情境改變而作出相應的行為調整。暫且不論倫理敏感度這種能力是否一定要有自我意識、心靈及反思等能力者才可擁有，但針對這種能力所認知或感受到的那些道德相關因素，並不排除仍有可能逐一分析及

羅列出來。當然這些因素或特徵可能因人、因情境、因不同事項等而有差異，相當複雜，但如果我們鎖定針對特定事項的倫理兩難決策，且兩難的內容涉及可能讓人失去生命，例如自駕車必須在有很高機率「強力撞擊行人」或「車上乘客將受到強力撞擊」之間做選擇（請注意，如前述這不等於說此為目標是撞擊兩方任一方，而不能是目標在減少總體傷亡人數，即使前述兩種都採取後果論的效益主義立場）、特定器官移植登錄系統在各條件組合起來分配順位不分軒輊的兩位病人之間做選擇等，諸如此類特定且重大事項的倫理兩難／衝突決策。

當我們將倫理兩難決策事項及範圍限縮許多之後，便盡可能去調查人們認為在這類決策中應該與不應該考慮的因素。而增加調查不應該的因素在於，研究團隊必須在那些被排除列入應該的眾多因素當中，進一步區分人們究竟認為不應該列入道德相關或只是認為無關緊要，後者可能無限多，前者往往觸及過去歷史上人們犯下歧視的錯誤。例如，自駕車兩難情境中，傷亡數量、有無遵守交通規則可能常被認為是應該列入考慮的道德相關因素，但種族、性別、年齡、社會階級、甚至車禍致死刑責與賠償金高低等，則常被認為絕對不應該列入考慮的因素，至於行人與乘客的外貌、體型等許多因素可能會被認為是道德無關。

此外，調查對象應該盡可能涵蓋這類決策的利益相關所有人員，即使沒辦法都涵蓋，那麼至少要符合社會科學量化調查研究的抽樣代表性等方法規範。而在調查這些人員有關這類決策中應該考量的因素時，還需詢問這些因素的優先考量順序，以瞭解對他們而言這些因素的重要程度與排序。不過人們或許不會被各因素的單一排序所困擾，但這些因素一旦有各種排序及程度或數量差異的組合時，可能就已經複雜到人們未必都能清楚說明為何從各種組合當中挑出其一。然而對於前述複雜性的掌握高低，或許正是對於個人倫理敏感度高低的考驗與評價：當個人所挑出的排序組合，也就是所做出的決策，受到越多利益相關人員的認同，可能就被認為倫理敏感度越高。

而利用機器學習的目的，不在於模擬人類的倫理敏感度，而在於模擬人類透過倫理敏感度所做出來的決策。就如同自然語言處理利用機器學習將某個語言翻譯成另一個語言，所模擬的是符合被翻譯出來的語言表達方式，其中包含該語言的語意、語法結構、甚至語用等，但不模擬這類雙語的人類翻譯者是如何翻譯的能力，而這種能力通常被視為評價個人是否擅長做某事的敏銳／敏感度。

所以，針對特定事項的倫理兩難決策所調查出來應該考量的因素，若能將它們列為輸入機器的資料特徵，並參考調查出來的各特徵排序來設定權重等，然後以機器學習找出那些特徵各種排序組合所產出的決策模型，再將決策結果徵詢該決策事項的利益相關人員，刪除不恰當並保留可接受的決策結果，不斷地回頭調整模型，直至模型所產出的決策結果在正確率也就是利益相關人員可接受程度達到一定數值時，便可視為其模型能穩定產出正確或合理決策的 AI 倫理決策系統。

所以就算接受倫理決策所需的倫理敏感度，並非以量化計算為本質的機器所能模擬，但機器可學習在特定事項中利益相關人員所曾做過的倫理兩難決策，可望最終能成功模擬人類做出那些決策的模式。但那項模式完全與人類做出類似決策的推論或感知方式[8] 不同，且往往不能套用在其他事項上，不若人類在具備一項能力後常能快速類推適用，以至於 AI 倫理決策系統至目前為止沒資格被稱為有倫理敏感度可言。簡言之，若以機器學習來設計針對特定事項的 AI 倫理決策系統，至目前為止 AI 都只是模擬人類在該事項上曾做過以及可能做出的倫理決策，可說只是道德殭屍（moral zombies）而已（Véliz, 2021）。

所以倡議設計針對特定事項的 AI 倫理決策系統，與其說是取代人類在特定事項上的倫理決策，倒不如說是人類透過 AI 來提升在特定事項上的倫理決策品質。人類應該有權拒絕由這類系統所提供的決策結果，但透過這類系統，期待降低有些人類決策已知常受到無關或不應該納入道德考量的因素所干擾之機率，例如，個人深思熟慮後可能也不認同的當下直覺反應、個人不自覺但得知後也不認同的隱性偏誤、決策生熟手的差異等。

---

8. 倫理推論究竟只包含理性的認知成分，還是也包含非認知的感知或情緒等，一直是後設倫理學中討論道德推論的構成要素之爭議重點（Richardson, 2018）。而這項爭論放在 AI 的脈絡中，作者相當贊同審查人的洞見：「在發展 AI 上經常提到的框架問題（the frame problem），正凸顯出人類就是因為有『情緒』才有辦法為有關行動的思考提供出框架，限縮行動的可能選項，使行動得以可能（De Sousa, 1990, Ch7）。相反地，少了情緒的 AI 因無框架而不知如何開始行動。換言之，情緒不見得對決策不好，甚至在某些情況下是不可或缺的。」儘管作者對於 AI 應否視為無框架有所遲疑，但即便 AI 有框架，那項框架是傳遞了系統開發人員擷取世界的某些面向而產生出的資料特性，而與人類是以如康德的理性克服情緒或如大衛.休謨的情緒指導理性的框架來採取行動，大相徑庭。所以作者才指出 AI 只是道德殭屍，所模擬的是人類外在的倫理決策行為而非內在的決策過程 （參考自：MIT Moral Machine website，https://www.moralmachine.net/）。

二、正確倫理決策無法從經驗資料獲得

不過對於機器不可能模擬人類倫理敏感度的駁斥，倚賴於以下預設：透過調查人類針對特定事項的經驗及看法，有可能獲知在該事項上人們認為正確、合理或可接受的倫理兩難決策。然而，若根本否定這類正確的倫理決策有可能透過調查獲得的假設呢？尤其涉及的是倫理兩難／衝突的重大決策，這意味決策結果本身就容易引起人們爭議，不容易達成共識，又如何確保透過調查就能獲知正確的決策，而不是只反映了受調者的觀點甚至偏誤而已？

（一）人工改良民主

以下將透過對於美國杜克大學研究團隊的實驗簡介，以回覆上述批評至少犯了兩項錯誤。該團隊指出目前器官移植分配決策遇到的困境是，或者負責分配決策的醫師在半夜臨時收到通知，或者相關決策委員會在無足夠時間了解移植資訊下，往往為了搶受贈器官移植時效，而被迫得在無法深思熟慮與評估的不佳決策狀態下倉促決定器官受贈名單。而他們認為如何分配器官移植才是公平的，除了依循相關移植法規及醫療專業人員的意見外，社會大眾的看法也應該納入考量。所以他們在設計有關腎臟移植分配的 AI 倫理決策系統時，採取四個步驟來將公眾意見納入（Sinnott-Armstrong & Skorburg, 2021）。

首先他們針對輸入 AI 的資料特徵選擇，以開放性的問題調查一般人、醫生和醫院管理人員在腎臟移植分配排序上，應該考慮與不應該考慮的因素各有哪些，以蒐集應該與不應該被列入機器學習的資料特徵清單。第二步驟是他們在編輯資料特徵時，除了將前述人員所提供的因素整併與清理外，還加入哲學家與倫理學家認為應該與不應該考慮的因素有哪些。

第三步驟則為了測試透過前二步驟所編列出來的資料特徵清單是否正確，所以他們調查不同於第一組的民眾對於清單的看法。而調查結果顯示，他們也如同前二步驟的人員認為，舉凡種族、性別、性傾向、宗教、政治立場、財富、接受社福補助款等，都不是分配腎臟移植順序時應該考量的因素，但諸如需接受移植的醫療急迫性、排隊等候時間、移植成功率、被診斷應接受移植前是否有抽菸、吸毒或酗酒等不良習慣等，則是分配腎臟移植順序時應該考量的因素。

最後步驟則是，研究團隊為解決應該考量的幾個因素在組合與排序上的兩難／衝突問題，便設計線上調查平台來尋求社會大眾的看法。平台上的兩難情境是包含兩位病人 A 與 B，在分別已知他們各自的年齡、有無或扶養幾位子女、診斷前每天喝幾杯酒等這三個在不同程度的組合條件下，請填答者選擇應該讓 A 或 B 先接受移植。選項中除了 A 或 B 可選擇外，還有丟硬幣這個選項來隨機決定誰應該先接受移植，目的是為了掌握填答者認為雙方在哪些三條件所構成的組合下其差異不大。

值得一提的是，選項不絕對二分，而是容許選擇有程度差異這種更貼近人類決策實際情況的調查設計方式，在清大研究團隊所設計的調查平台也可看到。該團隊將二選一改為五選一的選項，即除了贊成與反對之外，還多加了傾向贊成、傾向反對、無意見等三種選項，以期在輸入資料供機器學習時，能更精準模擬人類決策的樣貌：相同決策結果對於不同人來說，可能有決策信心程度的差異，而這種差異可能會與其他特徵的強度或先後順序改變有所關聯；有時候難以做出決策，則是因為兩邊的幾項關鍵特徵加總起來，可能都不足以構成讓人做出誰應優先被分配的決策。

在腎臟移植分配可能遇到倫理兩難／價值衝突這件事情上，杜克大學團隊透過綜合一般人與領域專家意見等多輪評估倫理決策因素的研究設計方法，以企圖找出人們認為最應該考量的道德特徵組合及正確決策模式，但批評者認為該研究團隊的做法不可行，由於無法跨越早在十八世紀哲學家大衛.休謨就提出實然與應然的鴻溝：從人們做決策的現實情境中，並無法找到人們應該如何做決策的規範（Braun, Broestl, Chou & Vandersluis, 2021）。

批評者以種族特徵尤其是美國黑人的研究證據指出，杜克大學雖獲知幾乎所有調查者都認為種族不應列入分配考量，但這樣反而如本文前述會產生種族盲的效果：目前全國各資料庫不論是腎臟捐贈者還是接受移植者的人數，在白人與黑人都相差懸殊下，這使得機器學習的訓練資料如依據這些名單來建立決策模型，不在資料特徵上特別標註種族，那麼即便調整醫療與社會相關健康因素的權重，所建立的模型其產出結果，仍然會讓黑人比起白人更難獲得腎臟移植。

然而類似上述批評至少犯了二項錯誤。第一項錯誤是，休謨所提出的實然與應然鴻溝儘管多年來受到不少挑戰，但即便在接受這項鴻溝下，人們提出有關某項決策應該或不應該列入哪些考慮因素的意見時，此時這些意見雖然屬於經驗資料，但卻是

評價性的經驗資料，而屬於應然並非實然的範圍，人們不只是單純在描述一項事實而已。所以研究團隊企圖從經驗資料中找出與腎臟分配決策相關的規範特徵，並不適用實然與應然鴻溝的問題。

此外，蒐集大眾意見儘管或無可避免包含集體偏誤，但研究團隊如果盡力透過研究方法的設計來去除，像杜克大學與清華大學在蒐集訓練資料時，除了調查一般人意見外，都會調查相關領域的專家意見等。尤其杜克大學還將整理自一般人與專家都認為應該與不應該納入決策考量的清單，再拿去給前述以外的其他一組人員再次詢問是否能接受該清單，並且針對選擇兩難／衝突的特徵建立一些假設情境來詢問大眾意見，例如建置腎臟移植分配的兩難決策眾包平台，然後再將最後都確定下來的特徵清單輸入機器學習，以便建立相關決策模型，最後請決策事項的利益相關人員協助挑選最適切的模型，或者依據個人的價值偏好與倫理信仰來客製化適合個人或特定群體的決策模型。甚至，可以將不同群體的決策模型拿來比較差異，或提供個人參考其他決策模型與個人偏好決策模型有何差異，當然也可回頭糾正相關模型的錯誤與偏誤等（Sinnott-Armstrong & Skorburg, 2021）。

杜克大學研究團隊將上述設計 AI 倫理決策系統的方式，視為「人工改良民主」（artificial improved democracy，簡稱 AID）。之所以是民主的，在於開發這類系統是建立在對於公眾意見的調查基礎上；而之所以是人工的，在於這種改良民主的方式是透過電腦程式來實現。他們倡議以 AID 來開發 AI 系統，將能廣泛應用在諸如醫療、法律、軍事、商業、甚至是個人生活等領域。換言之，從輸入機器的資料特徵選擇、清理、編輯、測試、分析、訂定，再到相關模型的建立、挑選、部署與調整等，若能納入一般民眾或專家知識的調查與參與，而不僅限於系統開發團隊，這樣所開發出來的 AI 倫理決策系統將有助人們了解或參考在相關決策事項中，那些經過深思熟慮與眾人智能淬煉出來的決策有哪些，或是了解自身與他人或其他群體偏好的決策有哪些差異，同時也可校正那些開發疏忽或混亂等有演算法偏誤的模型（Sinnott-Armstrong & Skorburg, 2021）。

AI 倫理決策系統依據杜克大學研究團隊所倡議的 AID 來開發，是否如他們上述所宣稱的，真有助於改善民主社會的某些決策，還是會削弱人類的倫理決策能力，尚待在現實世界的時間驗證。但透過黑人腎臟移植分配數據偏誤的案例之批評，確實有

一件事值得我們留意。

過往人類做決策常犯的盲點是，誤以為決策當下未納入不應該考量的因素如種族，最後所得出來的決策結果就是正確或公正的，而忽略了程序正義通常不保證結果的實質正義（許漢，2021）。雖然不確定杜克大學研究團隊所建構的模型在實際部署時，是否會拿有標記類似他們所蒐集的那些不應該納入決策考慮因素的資料庫，以作為相關模型產出結果是否符合公平性的檢測，而無法斷定該團隊所開發的系統未來是否會發生種族盲效應。但機器學習由於有資料驅動的特性，以至於不論相關開發團隊事先在選擇資料特徵等特徵工程階段，以及建立與調整模型等開發階段，究竟有多麼符合科學方法中減少偏誤的設計標準，並且在選擇特徵與解決特徵衝突時都納入公共民主參與程序，然而一旦在輸入現實世界真實資料的部署階段，就須格外小心從 AI 系統開發目標來看，所輸入的那些資料有無人口學的代表性，或是有無既存的人口學偏誤等。

諸如人口學上無代表性或有偏誤的資料，最容易反映在公開或官方所紀錄的資料上：通常不是紀錄太少，如美國黑人的器官捐贈與移植名單；就是沒有紀錄，如每年美國有多少人因犯罪紀錄而被排除在公共住宅等候名單；但有時候又紀錄過多，如警政系統的犯罪資料庫裡男性黑人就高的不成比例。而這些真實資料的缺陷或偏誤，往往就導致系統在學習時有所偏誤，最終所輸出結果就有算法或編碼偏誤（algorithmic or coded bias），亦即 AI 系統性的錯誤導致其應用結果，或者讓特定群體處於難以在道德上被合理說明的劣勢，或者讓特定群體在分類或預測上總是錯誤率較高，而使得這類 AI 被視為是一種會產生不公平結果的數位科技系統。

當然上述可以爭議的是，在評價 AI 系統是否符合程序正義時，所謂程序是指，AI 在部署之前包含目標設定、問題形塑、資料萃取、資料分析、資料預處理及特徵工程的系統設計階段，還是也包含模型的選擇、訓練、測試、驗證、修正等系統發展與部署階段。但不論程序正義與否的評價範圍到哪個階段，都可確定 AI 系統被實際應用的後果，究竟能否為社會所接受，而不會視為對特定群體有不公平對待之嫌，相關風險的掌握與預期，確實無法事先就能透過完善的 AI 系統開發與部署流程就能獲得。

（二）倫理兩難決策的公共參與基礎

最後，批評杜克大學研究成果者至少還犯了另一項錯誤是，當他們批評正確的倫理決策無法從經驗證據或透過量化調查來找尋時，往往缺乏積極說明，究竟如何找到正確的倫理決策？又如何定義正確的倫理決策？尤其本文所討論的 AI 倫理決策系統是有關倫理兩難的決策，這意味著相關決策必須從人們所珍視的重要價值彼此衝突下做出權衡，所以權衡結果，很可能既不一定有明顯對錯，也可能因人判斷而異。例如，沒有腎臟移植急迫性且剛登記等待移植的 A，以及有醫療急迫性且等待時間已有 3 年的 B，醫院此時如讓 A 比 B 優先接受移植，多數人會認為是錯誤的決策；但如果將 A 改為與 B 的醫療急迫性同等級，且等待時間 2 年又 10 個月，再加上還有 2 個未成年小孩需要扶養，醫院此時如讓 A 比 B 優先接受移植，或許不少人會同意，甚至認為是正確的決策。

既然倫理兩難的決策結果，無論選擇哪一方都未必有明顯對錯，還可能因人而異，那麼當決策事項將會影響個人權益或涉及公共事務時，在民主社會中誰有特權（privilege）而得以主張：像這類涉及價值多元與衝突的事項應該如何決策才是正確的？以及所謂正確決策的標準，是依照他們所信仰的價值觀或支持的倫理理論？

當特定決策事項涉及公共事務或與許多人的生活有關時，該決策事項的領域專家可協助提供決策所需資訊，並協助釐清決策的疑惑與概念，但最終應該要有該決策事項的利益相關人員參與，可說已經是當代自由民主社會相關公共政策實施前的基本要求與做法，同理也適用在開發針對特定事項的倫理兩難 AI 決策系統上。

前述已說明有必要研發這類決策系統，是因為已知人們對某些事項的決策，往往在時空或能力等限制下得匆忙或直覺反應，來不及深思熟慮，以至於所做出的決策結果或行為，是否與當下情境適切或合宜，不但常有運氣成分，也未必與決策者個人經過深思熟慮後的決策結果一致。而為了提升人類在這類事項上的決策品質，機器學習既然擅於從大數據中分析與找尋資料特徵之間的關聯模式，就可事先將考量特定事項應該相關的倫理特徵輸入機器當中，好讓機器學會如何產出一組適切的倫理決策模式，屆時就可利用 AI 來提升相關事項的決策品質，改善個人不及深思熟慮而作出事後可能都不認同的決策之缺點。

不過，機器學習應該事先輸入考量特定事項的哪些道德相關資料特徵，並讓機器

學會可產出哪些結果的決策模式，便是關鍵所在。同樣地，領域專家知識對於資料特徵及決策模型的選擇有幫助，但不應該是唯一來源。根據當代民主社會的基本理念，有可能受模型決策結果所影響的利益相關人員，甚至一般民眾，都應該有機會在 AI 系統週期的不同過程中被納入參與。像杜克大學便是讓利益相關人員與一般民眾有機會參與資料特徵的選擇，而後續的決策模型挑選、模型部署後的意見回饋等，按理說也應該有參與的機會與管道。

此外，如何避免 AI 複製人類社會既有的不平等，舉凡從資料的蒐集、清理與特徵選擇，到模型的建立、檢證與調整，再到模型部署的環境脈絡確認與限制，最後再到產出結果的監控與回頭調整模型設計甚至所輸入的資料等，亦即從 AI 開發到部署整個系統週期（lifecycle），便需要建立一套包含輸入資料、參數設置、歷次模型版本的健全管理方案。這項管理方案除了包含相關技術作業的設計與記錄之外，應該還包含對於 AI 系統目標的倫理價值承諾，那些承諾由於將表現在輸入機器的資料特徵之評估、篩選、建模與產出結果等面向上，若能採取平等主流化，將可以減緩或消除AI應用可能帶來不平等對待的後果。

伍、結論

本文主張使用 AI 來做人類所不擅長的倫理決策事項，並非認為 AI 可直接取代人類在這類事項上的決策，而是認為人類有可能借助 AI 找到對於這類事項的較佳或最佳倫理決策，進而得以改善自身在這類事項上的決策品質。

之所以期待 AI 在這類緊急且涉及倫理衝突的事項上將產出各種可能的較佳倫理決策，主要是人類往往受到如時間限制、環境壓力、資訊限制等客觀因素影響，或受到如疲倦、偏見、遺忘、緊張、疏忽、注意不集中等主觀因素影響，而難以作出較佳決策，或是所做出的倫理決策與未受主客觀因素影響下所做出的決策很可能不同。針對這類特定事項，若能事先蒐集人們做相關決策所會考慮的因素，那些因素是人們在盡力不受前述主客觀因素影響下做出深思熟慮的決策所會考慮的。然後再將所蒐集到的那些因素作為輸入機器的資料特徵，讓機器學習找出由那些因素在排序與重要程度有著各種差異組合的決策模型，該模型所輸出的結果亦受到人們的檢視、評估

及確認，那麼便可由這類 AI 系統直接執行所生成的決策。或至少可作為人們自我鍛鍊，即使在前述主客觀因素下仍有可能做出較佳決策的練習。而這樣的倫理決策形成過程，可視為是一種從資料驅動（data-driven）到 AI 驅動（AI-driven）的決策過程（Colson, 2019）：這類 AI 系統最終所提供的決策，不只來自所輸入的資料，且包含 AI 先提供一系列針對某緊急事項的各種較佳倫理決策選項模型，並在人類利益相關決策者的檢視與挑選模型後，未來以該模型所產生的決策才能被視為是最佳倫理決策。

不過上述願景要能實現，首要就是得開發出針對特定事項能產出可能是較佳倫理決策的 AI 系統。這種針對倫理尤其是涉及倫理兩難的 AI 決策系統，目前都還在初期嘗試中，但可確定的是，若期待以機器學習來成功開發這類系統，除了相關軟硬體技術的討論與處理外，就不應該忽略系統開發與部署過程中，舉凡有關輸入機器的資料蒐集來源、資料特徵選擇、不同資料特徵組合的衝突處理、模型建立後的挑選、驗證、測試與修改等，都將涉及倫理目標的承諾與倫理價值／原則的選擇問題。

因此本文透過檢討對於這類倫理兩難決策系統相關研究的批評，並區分為屬於方法論還是本體論方面的討論，分別討論這些批評有哪些值得注意，但有哪些可能出於誤解。

方法論的批評是指，不當使用眾包資料以及不當選擇資料特徵等二項。本文以美國 MIT 大學與我國清華大學所做的研究作為討論案例指出，只要研究團隊在使用資料時，清楚掌握眾包資料的特性與限制，並了解線上調查題項對於填答者的引導或暗示，透過適當設計，這類 AI 倫理決策系統未必不能使用來自眾包的資料。而資料特徵的設定與選擇，本身就會受到決策目標及問題設定的影響，但絕不以人類生命作為蓄意傷害目標，也不以人類生命作為需要被權衡的項目，並滿足雙重效應的倫理兩難條件，以及遵守基本人權規範燈，有必要成為 AI 系統開發人員都應遵守的設計規範，如此才稱得上是倫理的設計。此外，從資料特徵選擇、到算式的參數設定、再到歷次模型版本紀錄等，需要有一套健全的資料管理方案，方案中應該內含對於特定倫理價值／原則的設定。本文建議，平等主流化是值得考慮的選項與策略擬定。

至於有關倫理兩難決策系統相關研究的本體論批評，則是指計算系統無法模擬人類的倫理敏感度，以及正確倫理決策無法從經驗資料獲得等二項。本文區分模擬人類的倫理敏感度，以及模擬人類基於倫理敏感度所做出的決策，並指出 AI 倫理決策系

統可以僅就後者而不包含前者作為開發目標。簡言之，以此目標所開發出來的 AI 決策系統，充其量只是能做倫理決策但不足以視為有倫理決策能力的道德殭屍而已。

不過，這類系統如何產生出正確的倫理決策選項，便是關鍵。本文以介紹美國杜克大學的研究作為案例且援引其主張，有關如何設計得以產出正確的 AI 倫理決策系統本身，以及這類系統對於人類決策所帶來的影響，皆可視為人工改良民主的一種方法。這種方法本身就預設了，AI 系統從開發到部署，宜盡可能納入利益相關人員甚至一般民眾參與的重要。尤其倫理兩難這類決策本身，就涉及價值與倫理原則的權衡，不論是專家還是政府官員或是研發團隊等，皆無特權依照他們所信仰的價值觀來訂定何謂正確的倫理決策標準，可說是目前信奉自由主義的民主社會之基本信條。若我們仍相信這信條在 AI 科技來臨的時代也應該繼續適用，那麼有些事項若一定得在倫理兩難／衝突的情況下做出決策，不論有無利用 AI 來決策，所謂正確的、可接受的或是合理的決策，都應該是在盡可能尋求公共參與且獲得共識後的結果。

以上便是本文藉由檢討對於 AI 倫理決策系統相關研究的方法與本體論批評，以期釐清且主張，若有可能避免不當的方法設計，並對這類 AI 系統作出清楚的定位，人類有理由針對那些已知難以做出良好倫理決策的特定事項，嘗試開發 AI 倫理決策系統來改善或提升人們在緊急又涉及倫理衝突／兩難事項上的決策品質。

而這類 AI 系統所生成的倫理決策，有可能是模擬人類自認為在去除那些主客觀因素的理想情境下之倫理衝突／兩難決策，但更可能是協助人類找出如自駕車在道路上的兩難決策、器官移植的倫理衝突分配決策等諸如這類涉及公眾利益或公共政策的倫理決策。而尋找方式，就是確保這類 AI 在產出倫理決策的過程及結果，都應該符合民主社會的正當決策模式—利益相關人員甚至公眾的意見應該要被納入討論、審議及盡可能取得共識。這麼一來，利用 AI 的目的，不在於解決涉及公眾利益的倫理兩難／衝突，因為它們在價值多元的民主社會本來就是常態，而是借助 AI 讓人類不但能更精緻地處理這類倫理價值衝突的窘境，且得以找出更符合民主社會成員所期待的決策模式及決策結果。

參考文獻

中央社（2021.04.30）。〈台中嘉陽高中女學生校門前車禍喪命 肇事司機赴靈堂道歉〉。取自 https://www.cna.com.tw/news/asoc/202104300259.aspx 檢索日期 2023 年 7 月 21 日

王道維（2021）。〈文字標註與偏見處理〉，《人文社會 AI 導論線上課程第九集》，清華大學。取自 https://nthuhssai.site.nthu.edu.tw/p/406-1535-212970,r9286.php

甘偵蓉、許漢（2020）。〈AI 倫理的兩面性初探：人類研發 AI 的倫理道德與 AI 的倫理規範〉，《歐美研究》季刊，50 卷，2 期，231-292。DOI: 10.7015/JEAS.202006_50(2).0005

祖旭華（2020）。〈自駕車道德難題與問卷調查的研究方法〉，台灣人工智慧行動網。取自 https://ai.iias.sinica.edu.tw/self-driving-car-survey/ 檢索日期 2023 年 7 月 21 日

許漢（2020）。〈正義〉，《華文哲學百科》（2021 版本），王一奇（編）。取自 http://mephilosophy.ccu.edu.tw/entry.php?entry_name 檢索日期 2023 年 7 月 21 日

陳瑞麟（2020）。〈科技風險與倫理評價：以科技風險倫理來評估台灣基改生物與人工智能的社會爭議〉，《科技、醫療與社會》，30：13-65。DOI: 10.6464/TJSSTM.202004_(30).0001

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。DOI: 10.29843/JCCIS.202001_(38).0006

劉湘瑤、張震興、張瓅勻、趙恩、李思賢（2021）。〈人工智能倫理的挑戰與反思：文獻分析〉，《資訊社會研究》，41：27-64。DOI: 10.29843/JCCIS.202107_(41).0003

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., Rahwan, I. (2018). The moral machine experiment. *Nature*, 563, 59–64. Retrieved July 21, 2023, from https://doi.org/10.1038/s41586-018-0637-6

Bigman, Y. E., & Gray, K. (2020). Life and death decisions of autonomous vehicles. *Nature*, 7797, 579: E1-E2. Retrieved July 21, 2023, from https://doi.org/10.1038/s41586-020-1987-4

Colson, E. (2019). What AI-driven decision making looks like. *Harvard Business Review*. Retrieved July 21, 2023, from https://hbr.org/2019/07/what-ai-driven-decision-making-looks-like

Braun, E., Broestl, N., Chou, D., & Vandersluis, R. (2021). The challenges of using machine learning for organ allocation. Reply to Sinnott-Armstrong and Skorburg. *Journal of Practical Ethics*. October 15, 2021. Retrieved July 21, 2023, from https://journals.publishing.umich.edu/jpe/news/14/

Cervantes, J. A., López, S., Rodríguez, L. F., Cervantes, S., Cervantes, F., & Ramos, F. (2020). Artificial moral agents: A survey of the current status. *Science and Engineering Ethics*, 26(2): 501-532. https://doi.org/10.1007/s11948-019-00151-x

De Sousa, R. (1979). The rationality of emotions. *Dialogue: Canadian Philosophical Review/Revue Canadienne de Philosophie*, 18(1), 41-63.

Dixon, B. (2020/3/11). The "moral machine" is bad news for AI ethics. *Mind Matter News*. Retrieved July 21, 2023, from https://mindmatters.ai/2020/03/the-moral-machine-is-bad-news-for-ai-ethics/#

Fjeld, J., Achten N., Hilligoss H., Nagy A., & Srikumar M. Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI. *Berkman Klein Center for Internet & Society*, 2020. Retrieved July 21, 2023, from https://dash.harvard.edu/bitstream/handle/1/42160420/HLS%20White%20Paper%20Final_v3.pdf?sequence=1&isAllowed=y

Furey, H., & Hill, S. (2021). MIT's moral machine project is a psychological roadblock to self-driving cars. *AI and Ethics*, 2, 1: 151-155. DOI: https://doi.org/10.1007/s43681-020-00018-z

Gebru, T. (2020). Race and gender. In: Dubber, M. D., Pasquale, F. & Das, S. (eds.) *The Oxford Handbook of Ethics of AI*. Oxford: Oxford University Press. Retrieved July 21,

2023, from https://doi.org/10.1093/oxfordhb/9780190067397.013.16

Iagnemma, Karl. (2018) Why we have the ethics of self-driving cars all wrong. *World Economic Forum Annual Meeting*. Retrieved July 21, 2023, from https://medium.com/world-economic-forum/why-we-have-the-ethics-of-self-driving-cars-all-wrong-92566f282733

Jaques, A. E. (2019). Why the moral machine is a monster? *University of Miami School of Law*, 10, 1-10

Kalra, N. & Groves, D. G. (2017). The enemy of good: estimating the cost of waiting for nearly perfect automated vehicles. *RAND Corporation*. Retrieved July 21, 2023, from https://www.rand.org/pubs/research_reports/RR2150.html

Kauppinen, A. (2002). Moral Sentimentalism. In Edward N. Zalta (Ed.), *Stanford Encyclopedia of Philosophy*. Retrieved July 21, 2023, from https://plato. stanford.edu/archives/win2017/entries/experimental-philosophy/

Knobe, J. & Nichols, S. (2017). "Experimental Philosophy," *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.). Retrieved July 21, 2023, from https://plato.stanford.edu/archives/win2017/entries/experimental-philosophy/

McIntyre, A. (2019). The Doctrine of Double Effect. In Edward N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Retrieved July 21, 2023, from https://plato. stanford.edu/archives/spr2019/entries/double-effect/

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, ID., Gebru, T., (2019). Model cards for codel reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19). *Association for Computing Machinery*, New York, NY, USA, 220–229. DOI: https://doi.org/10.1145/3287560.3287596

Richardson, Henry S. (2018). Moral Reasoning. *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.). Retrieved July 21, 2023, from https://plato.stanford.edu/archives/fall2018/entries/reasoning-moral/

Sinnott-Armstrong, W., & Skorburg, J. A. (2021). How AI can AID bioethics. Journal of

Practical Ethics, 9(1). DOI: https://doi.org/10.3998/jpe.1175

Vallor, S., Ager, S., & Luan, R. (2021). The digital basanos: AI and the virtue of and violence of truth-telling. *In 2021 IEEE International Symposium on Technology and Society (ISTAS)*. DOI: 10.1109/ISTAS52410.2021.9629137 Retrieved July 21, 2023, from https://ieeexplore.ieee.org/xpl/conhome/9628888/proceeding

van Wynsberghe, A., & Robbins, S. (2019). Critiquing the reasons for making artificial moral agents. *Science and engineering ethics*, 25, 719-735. DOI: https://doi.org/10.1007/s11948-018- 0030-8

Véliz, C. (2021). Moral zombies: why algorithms are not moral agents. *AI & SOCIETY*, 36(2), 487-497. DOI: https://doi.org/10.1007/s00146-021-01189-x