Occasional Paper No. 244 *March 2021*

被害者視角
刑法何以保護人工智能體？

賈健

香港亞太研究所

Hong Kong Institute of Asia-Pacific Studies
The Chinese University of Hong Kong
Shatin, New Territories
Hong Kong

被害者視角
刑法何以保護人工智能體？

賈健

香港中文大學
香港亞太研究所

引用本文
賈健。2021。《被害者視角：刑法何以保護人工智能體？》。取自香港中文大學香港亞太研究所網站：http://www.hkiaps.cuhk.edu.hk/wd/ni/20210308-105600_3_op244_t.pdf

香港亞太研究所 • 研究專論第 244 號

作者簡介

賈健，法學博士，西南政法大學法學院副教授，碩士生導師，重慶大學法學院博士後；研究方向為刑法哲學。

鳴謝

本文是 2018 年度最高人民法院司法研究重大課題「刑事裁判公眾認同問題研究」（批准號：ZGFYKT2018-1905）；2018 年度重慶市教育委員會人文社會科學研究一般項目「人工智能體的刑法歸責問題研究」（批准號：18SKGH007）。論文的部分研究工作在香港中文大學香港亞太研究所訪問期間完成。

ISBN 978-962-441-244-4

被害者視角
刑法何以保護人工智能體？

《西部世界》之殤：問題的提出

當下人工智能機器人（簡稱智能機器人）正以迅猛的速度進入我們的世界，尤其是那些如自動駕駛機器人、手術機器人、性愛機器人、情感慰藉機器人等會對我們社會生活產生巨大改變的機器人，已經引起了人類社會廣泛的關注和重視。可以說，智能機器人的出現及其類型與數量的劇增，某種程度上改變了傳統的社會關係，並加劇了我們這個時代下社會生活的不確定性。具言之，它們除了帶來前所未有的便利性以外，從目前看，也帶來了一些新的危害，這引發了人們的普遍不安，進而科技界、哲學界與法學界等領域的學者開始思考人工智能體侵害人類社會的防範與規制問題。但其實，人工智能體作為新生的社會存在物，隨着不斷地與人類社會深入互動，其不但會具有犯罪者的側面，同時也會具有被害者的側面。後一種面向以及其帶來的一系列法律問題，同樣值得學術界關注和反思。

2016 年美國 HBO 電視頻道首播的《西部世界》（Westworld）科幻類連續劇就向我們展示了智能機器人作為被害者，最終產生自我意識後的情景。該劇演繹了在未來

世界裏，人類創造了上千與人類外形無差別的機器人「接待員」，在一個巨型的成人科技樂園——西部世界裏，他們重複着管理員預先設定的活動程序，配合服務到此的人類遊客。懷着各種動機的人們花錢進入這個樂園，並在這裏奸淫、擄掠、殺戮，滿足自己在現實生活中無法實現的欲望，「接待員」們則在其中扮演一個個被挑釁、受凌辱和被槍殺的角色，這些「接待員」不僅具有超高的仿真外形，還具有自身的情感，能夠帶給遊客最真實的體驗。比如，其中彈以後會流血，受傷以後會痛苦地哀嚎等。待夜幕降臨後，所有機器人的記憶被清除歸零，等待第二天新一批入園的遊客。最終機器人「接待員」們在無數次被虐殺的相似情境中擁有了自我意識和思維，開始了一場對人類的瘋狂報復……應該說，《西部世界》所反映的並不是對人工智能體發展前景的危言聳聽，而是對未來人類與智能機器人關係的一場深刻反思。

對此，我們有必要思考一個機器人技術發展的終極社會問題，即能否將具有自主學習能力的智能機器人當作與人類一樣的被害主體來平等對待？對智能機器人的某些非道德性奴役應否制止並使之得到刑法的平等保護？如果答案是肯定的，那麼智能機器人也將成為未來刑事犯罪的被害人，此時，我們又該如何救濟其被犯罪行為所侵害的利益呢？下文將圍繞這些問題展開分析。

人工智能體應具有法律上的主體資格

就上述《西部世界》所引發的一系列問題而言，其有一個共同的前提，即智能機器人是否具有法律上的主體資格。傳統理論認為，法律上享有主體資格者是指享受權利和承擔

義務，且一般情況下具備權利能力和行為能力者，通常是在人和由人組成的群體範疇下探討，包括由法律擬製的法人。這一傳統的人本主義視角決定了在制定法律時，大多數人會將那些非人類的存在物排除在主體範疇之外。但是，這種以人類為中心的法律主體資格之立場正在受到理論與現實的雙重挑戰。例如，隨着對生態環境法律保護研究的深入，愈來愈多學者開始提倡應賦予包括動物、植物、環境、自然和生態系統等非人類存在物的法律主體資格。[1] 一些國家的立法也為某些非人類存在物在生存或存在權上提供了法律保護；[2] 美國的一些司法判決甚至讓鳥、猫和狗成為了訴訟中的原告或被告。[3] 正如有學者所指出的，法律主體資格具有歷史性和開放性，其範圍會隨着歷史的變遷不斷擴大，尤其是動物、法人權利主體地位的獲得，説明了「物種差異不再視為獲取

1. 如有學者認為，人們已經在道德上承認了享有主體資格的主體不僅限於人類，還包括動植物、環境和生態系統，承認它們具有存在的權利和內在價值，那麼法律也應該對此給予保護，因為非人類存在物獲得要求正義的資格（參見曹明德，2002:117）。

2. 美國伊利諾伊州的《人道地照料動物的法律》（*Humane Care for Animals Act*）規定：動物養育者必須為動物提供足量的、質量好的、適合衛生的食物和水；充分的庇護場所和保護，使其免受惡劣天氣之害；人道的照料和待遇。禁止任何人打、殘酷對待、折磨、超載、過度勞作或用其他方式虐待任何動物（參見江山，2000:30–31）。

3. 1979 年，美國聯邦法官 Samuel King 為保護生活在夏威夷州的帕里拉（Palila）屬鳥作出了判決：夏威夷當局被要求必須在兩年內完成禁止在毛納基火山（Mauna Kea）放牧的工作（見 *Palila v. Hawaii Department of Land and Natural Resources*；另參見曹明德，2007:163）。

權利主體地位的法律障礙」（張玉潔 , 2017:58）。本文認為，這些探討實際上為智能機器人獲得法律上的主體資格掃清了部分前提性的障礙，進而言之，其法律主體資格的存在有其內在必然性。

人工智能體應當享有法益

首先，智能機器人有值得法律保護的利益，即智能機器人應當享有法益。之所以沒有用「權利」替代「法益」，是因為法益是權利表達的內容，而權利只是表達法益的工具，但不應當是唯一的工具。權利往往直接與人相對應，並非所有關係領域的法益都適合用權利的概念來表達，「勉強的不加改造地將權利模式移植到國家統治及人與自然關係領域，創設國家權利、動物權利、大自然權利等概念，並試圖借助於原有的人的權利的分析模式去解釋上述所謂權利，實際上忽視了權利的工具價值的有限性，過分注重了權利的價值性表現，在現實中會遇到重重阻力。」（焦艷鵬 , 2012:20）

同樣，認為「智能機器人擁有權利」的觀點會受到「智能機器人並不等同於人」的質疑，而去掉具有工具價值的權利概念的外觀，智能機器人完全可以擁有權利要表達的法益內核。Isaac Asimov 的機器人三原則（Three Laws of Robotics）得到了廣泛的認同，而第三法則是，在不違背第一法則及第二法則的情況下，機器人必須保護自己（阿西莫夫 , 2005:273），表明了機器人應該享有類似於人的生存權以維持其自身存在的利益。這種維持自身存在的利益需求既是機器人自身最基本的倫理要求，是智能機器人發展的起點，其實也是人類社會得以發展的基本要求，畢竟保障機器人不受無端破壞，才能實現將其用於社會生產服務，提升人

類福利的最終目的，即這種存在利益同時體現了道德性與功利性目的。

而且，智能機器人也應該享有一定程度的活動或選擇自由的利益，保護智能機器人的此種利益是推動智能機器人技術發展的因素，也是推動其與人類建立友好合作關係的必要步驟。Alan Turing 認為，真正的人工智能並非是要超越人類思維，而是能製造出與人類一樣思維的智能。所以他認為：「如果一台計算機的行為方式與人類一樣，那麼就可以說它是智能的。」（托比 • 沃爾什 , 2018:33）

很明顯，就目前人工智能體技術的發展趨勢來看，其已經超出了在程序性操作下的規行矩步，開始了在大數據的支持下進行深度學習和深度推理的進程，「機器可以根據明確編碼的知識進行推理，或是依靠與現實世界的互動來學習」（托比 • 沃爾什 , 2018:50），最終，將會使智能機器人在與人類互動的過程中做出與人類相似的反應，從而代替人類進行某些行為。創作型機器人的出現，恰好證明當前的智能機器人必須具備一定的自主性，它們並不完全受編碼的操縱，具備同人類一樣的創造能力進而實現其價值。同時，要將如自動駕駛機器人一樣需要在複雜多變的環境下工作的智能機器人投入到社會生活中，就必須提高其自我判斷的能力，畢竟自動駕駛機器人面臨着極其複雜的交通狀況，需要應對諸如行人、十字路口、其他車輛臨時變道等意外因素，當遇到類似於著名的「電車難題」之兩難困境時，其應當能夠獨立作出行為判斷和選擇。

除此之外，還應該保護智能機器人獲取和保留數據資源的利益，這是使其保持智能化的基礎。人工智能的發展離不開大數據的運用，不論是深度學習還是推理，智能機器人的社會化應用都是建立在對大數據的採集、決策技術和算法的交互使用上（張玉潔 , 2017:62），智能機器人的發展深度，

便取決於對數據的可採性與公民的隱私權之間的矛盾解決，以及數據的量化保證上。

總之，智能機器人技術的研發和社會化運用要求承認並保護其某些內在的、能使其維繫其自身存在的特定利益，而在上述利益均得到基本滿足的前提下，智能機器人理應獲得法律上的主體之承認。換言之，智能機器人這一角色的存在及發展，本身就已經宣示其必須主體性地享有法益。

人工智能體具備利他性

其次，人工智能體具備利他性。傳統的法律義務是與法律權利相對應的概念，既然權利的內核是法律主體的自我利益訴求，具有利己傾向，那麼義務的本質，就是對其他權利主體利益訴求的滿足，就是一種利他性。從廣義上講，這種利他性是客觀的，既可以是對人類的良性表現，也可以是一種對包括人類在內的客觀環境的良性表現。不過法律義務是站在人類的角度，以行為為出發點來設計的，而如果將目光投向一切存在物，那麼這種利他性就是義務的另一種表達。按照這種觀點，世間萬物都具有利他性，但是只有通過價值權衡，並被法律規定了的部分，才具有法律上的意義。動物的利他性之所以受到法律規定，是因為動物是生態系統中的一個重要部分，具有維持生態平衡的利他性，人類也能從其不受無端驅趕和殺戮中獲得自身的生存利益，所以法律將動物的生存利益納入保護範疇。

智能機器人的利他性體現在機器人自身的存在價值對人類社會的直接意義。工業機器人正在以其無可比擬的優勢進入工廠，如今世界工業機器人製造商「四巨頭」之一的發那科公司（FANUC），因為沒有人力而不需要照明，而成為了著名的「黑暗工廠」（托比 • 沃爾什 , 2018:60）；自動駕

駛的人工智能系統正在降低人工成本，其依賴算法形成的最優路徑也能極大地減少交通擁堵，提高公路的利用率；在醫藥領域，人工智能則成為新藥篩選和安全性檢測的得力幫手，其利用策略網絡、評價網絡，以及蒙特卡洛樹搜索算法（Monte Carlo tree search），從萬千備選化合物中挑選出最具有安全性的化合物，大大地節約時間和成本（高奇琦，2018:104–05）。

從現實角度看，應該說，絕大多數智能機器人都設計於並實際服務於人類，為人類社會帶來效率和價值，可以說利他性是智能機器人存在的出發點。當然，由於利益之間也常常產生衝突，所以利他性應該立足於大多數人的利益，而非少數人或小團體的利益，尤其是對於那些基於不法意圖生產出來的單純破壞性機器人，因為不具有利他性而不能賦予其法律主體資格。

人工智能體具備可責性

最後，智能機器人具備可責性。智能機器人的利他性不完全等同於動物的利他性，應該說，智能機器人是更為貼近人類的一員，要與人類進行長足而深刻地交往，要進行一系列類似於人類的活動，必然要求其活動符合社會規則和遵守社會秩序，當它們的活動觸及規則底線的時候，就產生了責任追究。責任來源於對義務的違反，刑法上的責任要求主體對自己行為性質和後果具備認識能力和控制能力，即對罪過的要求。

智能機器人的罪過認定基礎，一是表現在其獨立判斷能力上。當前人工智能被分為弱人工智能和強人工智能，兩者的區分標準在於是否可以進行一定的獨立性判斷與決定。強人工智能被認為是能夠進行推理和解決問題的智能機器，它

像人一樣有知覺和意識（王肅之 , 2018:56）。智能機器人的自主學習和深度推理能力，可以使其在輸入數據之後分析潛在的規律，推算出新的結果，其實質已經部分脱離了人的控制，帶有了一定程度的判斷和行為獨立性，所以根據罪責自負原則，將來智能機器人犯罪不能完全歸咎於製造者或所有者。

二是智能機器人感知力的獲得。智能機器人的感知力借助於傳感器的應用，其中，計算機視覺成為多數智能機器人的一個重要組成部分，通過物體識別、運動分析和姿態（位置和方向）估計等其他通用任務，機器視物取得了很大進展。計算機在語言處理上取得的進步，也使智能機器人加深了對自然語言的理解和使用（托比 • 沃爾什 , 2018:62–66）。當前神經網絡技術、傳感器技術和語言技術的結合，有利於進一步增強人工智能對外界的理解和感知能力。加上現在多方研討要加強智能機器人的倫理道德建設，[4] 例如，美國機器人研究專家為了使軍用機器人比人類更具有人性，在智能機器人系統中設計了「人工良心」，並公開徵求智能機器人應遵循的道德規範（杜嚴勇 , 2014:100）。這表明賦予人工智能以人類的價值觀念是發展智能機器人技術的必然要求，即使智能機器人無法擁有同人類一般的道德理念，也要求其能夠在特定環境下作出符合人類價值的判斷和選擇。

另外，智能機器人從被研發投入使用之日起，就擬定其能夠從事可控的行為，加上智能機器人實體可以評估未知結

4. 例如，2017 年 1 月，在美國加利福尼亞州阿西洛馬（Asilomar）舉行的「向善的人工智能」（Beneficial AI）會議上，針對人工智能的未來以及監管問題，列出了一份有 23 條的原則列表，其中 13 條涉及人工智能的倫理和價值。

果與實行行為之間的發生概率（王耀彬 , 2019:142），相比於人類來說，在預知行為所引起的結果發生可能性上具備更為「天賦」的能力，在評估結果發生可能性的基礎上做出合理的行為，其實就是控制力的表現。當然如果因為設計、生產環節存在疏忽，或者有人刻意破壞智能機器人的控制系統，導致智能機器人缺乏控制能力而做出危害行為、產生危害結果，則只能歸責於製造者或破壞者而非智能機器人，因為某種程度上說，不可控的智能機器人與精神病人一樣缺乏刑事責任能力。因此，只要智能機器人設計者和製造者盡到了必要的技術注意義務，智能機器人在感知系統幫助下，可以正常分辨事務和控制自身行動，而且能夠獨立做出行為時，它所造成的客觀危害就應該由其自身承擔法律責任。

再者，智能機器人固然不能像人類一樣承擔肉體上的痛苦，但是讓智能機器人受到懲罰或彌補的機制早已有探討，比如為將來人工智能主體設置「資格刑」，以防止其再犯罪（王肅之 , 2018:61）；或者考慮建立賠償基金，作為強制保險制度的一個補充等（司曉、曹建峰 , 2017:172）。甚至還有國外學者提出，每個智能機器人都是超越知覺的綜合體，其能力具有二次性，並非自然而生，但真實地具有（一定的）能力，如果一個自主或者部分自主的智能機器人犯罪，不能將其歸因於自然人或者法人，而應將之視為是其終身的恥辱，輕者應斷電一周（Gleß and Weigend, 2014:577–78）。

由此，智能機器人獨特的利益訴求、與生俱來的利他性，和可責性的實現可能性，使其具有成為法律乃至刑法主體的內在必然性，這說明其正在或將會逐漸脫離客體的藩籬，隨着人工智能技術的發展和社會倫理認知的深化，智能機器人將完全可能成為法律主體的一員，而不再是被視為單純受奴役的物質性工具。這種社會地位的變化恰恰印證了歷史上的奴隸、有色人種、動物，再到被歧視的女性獲得法律主體地

位的演變過程。對此，有學者提出反對意見，認為「奴隸、婦女、黑人、動物、法人獲得權利主體地位，都以人的活動為基礎並以人類的安全和福祉為前提，人工智能與之類屬不同，不具可比性，不應被賦予獨立的權利主體地位和承認其獨立的利益。否則，當智能機器的利益與人類利益衝突時，未來遠比人類智慧強大的、具有自主意識的超級智能必然全面碾壓人類反抗，使人類處於被奴役甚至滅絕的境地。」（皮勇 , 2018:152）然而，這觀點似乎有違歷史辯證法的方法論，在諸如奴隸、黑人被歧視的年代，站在奴隸主和白人的所謂「主人」立場上，並不會覺得賦予奴隸、黑人法律主體地位會有助於他們的安全和福祉，實際上，對於這些主體身分的賦予，並不取決於當時佔據「主人」地位的群體之認知與價值判斷，而是具有歷史發展的客觀辯證性。

其實上述論者仍是閾於人類中心主義的立場來考慮問題的，保證人工智能體不危害人類，與是否在一定的前提條件下賦予其法律主體地位，並不存在必然的衝突，我們完全可以在協商制定人工智能體倫理規則與保障規則的基礎上，賦予其一定的法律主體地位，從歷史發展與人類整體角度看，這並不違背人類的利益。

刑法應賦予人工智能體以獨立的被害者地位

人工智能體承載着人類的基本道德情感

從刑事立法角度看，刑法是由統治階級將那些嚴重悖離社會道德、違反社會秩序的行為規定為犯罪的法律。Joel Feinberg 將損害原則和冒犯原則作為刑法犯罪化的完整道德基礎，損害原則被認為是對人類福利性利益的破壞或妨礙，

而人類在各個階段的生理健康、精神狀態上的善好利益，以及人生的遠大抱負，都是人類應該享有的福利性利益；而冒犯行為是引起他人不快的精神狀態的行為，其具有導致損害結果的可能性，刑法只能對極其嚴重，且受眾難以避免的冒犯行為進行規制（鄭玉雙 , 2016:185–86）。

冒犯行為因為直接與滋擾行為引起的人們的精神不快結合起來而被認為與情感有關，但本文認為，其實損害原則所依據的福利性利益，歸根結柢也是情感的宣洩與表達。因為肉體的劇烈疼痛會伴隨精神的傷害，他人生命被非法剝奪，會給其家人朋友帶來痛苦與折磨，財產受侵奪、欺騙、被他人非法佔有，也會讓所有者因為失去物質保障而無助失落。某種意義上可以說，享有福利性利益的最終目的，是為了提升個體的精神狀態，使情感有所依托並得到釋放。其實，原始的人類社會規則就是建立在集體道德觀上的，行為一旦觸碰道德底線，就會遭受懲罰。正如英國社會人類學家 Alfred Radcliffe-Brown 所言，原始社會中，「一個社會中公認的不法行為……，其核心就是群體對因為內部成員侵犯了公認的群體道德觀念而導致的社會動蕩狀態的反應。在這個反應中包含了集體道德憤怒的情感，從而起到使社會恢復安寧的作用。它的最終目的就是保持社區成員的最基本的道德情感。」（A. R. 拉德克利夫 • 布朗 , 2014:191–92）

刑事新派代表人物 Raffaele Garofalo 也認為，犯罪不是對權利的侵害，而是對基本道德情感的侵害（加羅法洛 , 1996:44）。隨着理性的成長，人類社會才逐漸將懲罰的依據由集體道德外化為貌似更具客觀性的利益，但從根本而言並沒有完全擺脱集體道德觀念的影子。可以說，不法行為引起利益的客觀狀態面受損，只是犯罪化正當性的表面依據，基於利益歸根結柢是「人」的利益之理解，本文認為，犯罪化的實質正當性，仍在於抗制和打擊嚴重破壞社會成員情感的

行為。對此，梁根林（2005:41）曾言，道德的基礎，是將社會中的人假設為普遍善良的個體，並且以此作為社會治理的重要標準，以期社會中的普通人都能成為無我、忘我的天使，而刑法存在的基礎，是將社會中的個體假設為性惡的個體，利用刑法的治理，就是利用刑法壓制人性之惡。

未來，智能機器人不但將以深入社會的交往方式與人類進行情感交流，它們自身也完全可能產生與高級生命物一樣的感覺和情緒。Alan Turing 的圖靈測試（Turing Test）就強調了智能機器人除了要具備人類感知能力之外，也要具備能夠與人類進行情感互動的能力，「在重視人工智能完成任務和功能強化的同時更要建立和滿足人的情感和心理需求，這才是人工智能的最終定義。」（張愛萍 , 2016）

現實生活中也需要愈來愈多的能夠與人類進行情感交流的智能機器人，比如伴侶機器人，它需要能夠通過對人類面部表情、語言表達、肢體動作等外在表現進行情感計算和分析，從而讀懂人類情感並表達自己的想法，滿足我們對社交的內在渴望。為了實現人類的陪護需要，未來智能機器人也會相應地產生疼痛、難受、恐懼或快樂等現有的生命體所具有的感知力，以減少與人類的溝通障礙。當智能機器人也具有可以表達自己情緒的能力時，就有必要探討是否應該將人類對智能機器人的某些行為納入道德和法律的範疇。Ray Kurzweil 在《人工智能的未來：揭示人類思維的奧秘》（*How to Create a Mind: The Secret of Human Thought Revealed*）中說過：「當機器說出它們的感受和感知經驗，而我們相信它們所說的是真的時，它們就真正成了有意識的人。」而大多數道德和法律制度也是建立在保護意識體的生存，和防止意識體受到不必要的傷害的基礎上的（高奇琦 , 2018:29）。

與之相應，人類也會將重要的情感寄托在某些智能機器人身上。如日本的陪護機器人「帕羅」（PARO），可陪伴

老年人唱歌、跳舞、遊戲等，頗受歡迎。為了獲得更逼真的體驗效果，愈來愈多的性愛機器人無論從外形上還是從內在情感互動上都逐漸趨近於人類，尤其是矽膠打造的皮膚和面龐，能夠增加人類的親切感，這種與人類貼身接觸的機會，難免讓人類對其產生依戀、愉悅的情感。而如果這些基本的情感遭到破壞，就容易對這些破壞行為產生出違背道德觀念的罪惡感。2014 年，一位人機交互（human–computer interaction）專家做了一個實驗，她讓人類主動傷害形象真實可愛的機器人，然後記錄人類的身體反應。實驗後，大部分人類都表示這一行為讓他們感覺到深深的不安，道德意識較強的實驗者甚至產生了對於自我的較強的抵觸意識（搜狐，2016）。

從現實角度看，人類對智能機器的情感最初表現為依賴感，即將個人決策建立在智能機器的分析信息之上。比如當前我們依賴智能導航系統提供的駕車路線行駛，醫療診斷依據大數據提供的概率進行判斷，出行的意願會受到智能算法得出的天氣預報影響等。等到人工智能發展出情感互動等類人功能，人類原先基於機器人的決策功能產生的依賴感，會逐漸轉變為更深層次的依戀感，並進而延伸出諸如憐憫、同情等情感。[5] 事實上，憐憫或同情是人類產生的一種由己及他的良善情感，這種情感並不僅限於對人類或動物，只要是一種與人存在情感互動之物處於被侵害狀態，都有可能觸發

5. 打個也許不恰當的比喻，這就像幼童對於父母感受的變化那樣，一開始只是出於父母滿足了自己基本的生活需求而產生依賴感，但是父母不斷給予幼童情感的呵護，幼童就會逐漸將依賴感轉變為依戀感，即使幼童長大不再依賴父母，也會基於依戀感與父母保持長久而親密的關係，從而有了基本的家庭倫理道德。

人的憐憫、同情之心。智能機器人技術的發展，將使機器人從外形、思維方式和行為方式上都比動物更像人類，人類會更容易把智能機器人當作人來看（Scheutz, 2012:207），所以當智能機器人受到傷害時，人們會更容易產生憐憫和同情，從而讓人對這些傷害行為發出道德情感上的譴責。例如，美國軍方曾經讓智能機器人踩踏地雷進行拆除的測試，但最終上校下命令終止了該測試，原因是每當該智能機器人踩到一個地雷，它就失去一條腿，並借助剩下的腿繼續進行，「上校無法忍受看到被燒傷、傷痕纍纍和殘廢的機器用最後一條腿拖拽行動的痛苦，他控訴這項測試是不人道的。」（瑞恩•卡洛、邁克爾•弗魯姆金、伊恩•克爾 , 2018:220）

本文認為，這種移情正是將侵犯人工智能體的行為予以犯罪化的倫理根基。基於刑法的道德性所應該與能夠發揮作用的範圍並不排斥非自然生命物，因此，如果刑法關注的是犯罪行為本身及其對他人的影響的話，那麼，將對人類的犯罪行為施加到智能機器人身上，同樣也應該被認為是犯罪行為。從刑法上重視這種道德情感的原因，還在於刑法的人道並非「對人」之道，而是「為人」之道，殘酷本身除了會使民眾對刑罰變得「麻木不仁」，使刑法的預防目的落空之外，還會鬆綁人的道德約束，從而增加犯罪行為發生的概率。有論者指出，如果法律沒有對針對智能機器人的傷害虐待等行為進行保護，便極有可能為公眾帶來消極情緒，從而縱容人類的殘暴行為，並最終使這些行為轉化為針對他人的違法犯罪行為（劉憲權 , 2018:199）。總之，如果我們能夠接受甚至力促將智能機器人作為刑法上的犯罪人主體，就沒有理由否認其亦能夠成為刑法上的被害主體。

智能機器人已經嵌入了人類社會秩序之中

刑法作為國家的統治工具，具有維護社會秩序穩定的功

能，在社會結構的調整過程中，這種功能被內化，進而成為犯罪化的正當性依據之一。事實上，自 1997 年中國的新刑法頒布以來，立法者基於社會治理的需要，從實現社會安全、秩序穩定和經濟發展的立場出發，增加了很多反道德性並不明顯的行政犯（時延安 , 2018）。在風險刑法和預防刑法理論盛行的當下，法益的提前保護為秩序維護這一刑法的正當性根據提供了更多的施展空間。應該說，刑法維護社會秩序的目的轉向，使得對智能機器人的刑法保護更具可能性。原因在於智能機器人從某種程度上說，已經深深地嵌入了當下的社會秩序之中，承擔了各種重要的社會職能，一旦其受損，遭到侵害的實則是其背後的社會秩序與社會機能。

具言之，當前智能機器人通過充當社會各個領域的工作者和活動的參與者，已經與社會秩序產生實質性的緊密聯繫。智能機器人加入的社會領域已經愈來愈廣泛，高度智能化帶來的優勢使其不再局限於完成簡單低級的工作，而是正在成為某些領域的精英。比如，將殺傷性軍用機器人投入戰爭；自動駕駛機器人投入交通運營；智能寫作機器人開始投入新聞、股評、詩歌乃至小說的創作；醫療機器人投入醫用手術；家庭服務型機器人投入家政服務市場；在金融領域，智能機器人在速度和數據整合準確度上，已經逐漸超過金融分析師，現在紐約和倫敦證券交易所的交易大廳幾乎形同虛設，真正的交易過程已經全面實現了「機器自動化」（中國日報網 , 2016）。未來，智能機器人甚至可能走進政治領域，成為辦事高效的公務協助員，從事大量行政事務性工作，尤其是成為交通領域的機器人警察，運用電子眼提取並分析交通事故數據，極速開出罰單，提高交通事故處理效率，也能夠以智能定位系統追蹤可疑車輛，或者指示車輛行進暢通路段，大街小巷都可以看到能夠經受日曬雨淋的電子交通巡警機器人。

將來智能機器人大範圍參加社會各個領域的實踐，導致它們成為維繫社會秩序穩定和創造新秩序的重要一員。所以，未來對智能機器人的損毀、虐待、剽竊或竊取數據等行為，完全可能成為破壞社會秩序的嚴重行為。例如隨意損毀或毆打在行政機關從事公務的智能機器人，已經危及了政府的管理秩序；為了毀滅證據而消除記錄了刑事案件關鍵信息的智能機器人的數據，危及了司法秩序；剽竊智能機器人生成的具有獨創性特徵的智力成果，破壞了著作權秩序；竊取並控制智能機器人金融分析的重要數據，威脅了市場金融秩序等。

之所以要重視這些被智能機器人參與的社會秩序，原因在於這些領域承載了人們的重要利益，在數據共享時代，人類的重要利益一般都屬集體利益，而某一領域的秩序混亂，最終會使集體利益受到嚴重損害。此外，一個侵害智能機器人的不法行為除了導致社會秩序混亂，還會進一步引發其他犯罪現象。著名的「破窗理論」（broken windows theory）可以很好地揭示無序的環境對犯罪的影響。[6] 這裏的無序不單是指物理環境的髒亂差，還包括惡劣的人際關係及越軌行為。這表明無序對人的越軌行為或者違法犯罪產生了強烈的暗示性或者誘導性，因為無序體現了某種程度上犯罪控制力

6. 破窗理論認為，無序「將使一個社區以螺旋形的方式慢慢失去控制，其中的居民也會漸漸躲避、退出或者逃跑；這種結果又反過來進一步加劇了社區中非正式控制機制的消失，並導致更為嚴重的犯罪，進而導致恐懼的增加，等等。隨着社區的衰敗，無序、恐懼以及犯罪螺旋式上升。」參見麥克•馬圭爾等（2012:683）。

的薄弱（參見李偉 , 2014:137）。智能機器人全面嵌入社會領域後，任何一個破壞智能機器人正常活動的行為都有可能製造出社會的無序性，從而鼓勵其他違法犯罪現象的發生。比如，剽竊智能寫作機器人的智力成果的行為如果得不到法律的及時制止，就會讓潛在的犯罪人看到獲益的契機而肆無忌憚地剽竊智能機器人的智力成果，甚至會蠱惑更多守法公民剽竊他人的智力成果，最終擾亂著作權領域的秩序，削弱人們的創作熱情，還變相鼓勵了不勞而獲的非誠信行為，導致減損人的整體道德感，從而又引發其他的犯罪行為。

由於未來智能機器人承載着人類的基本道德情感，與社會秩序產生愈來愈密切的聯繫，所以刑法有必要進一步考慮對智能機器人的保護，使其免受不法侵害，這也是發展人工智能道德的必要內容。最新的歐盟委員會「可信賴的人工智能道德準則草案」（Draft Ethics Guidelines for Trustworthy AI）指出，「我們必須確保最大化 AI 的優勢，同時降低風險，因此需要以人為本的人工智能方法，AI 的開發和使用不應被視為一種手段，應視為增加人類福祉的目標。」（搜狐 , 2018）其將確保人工智能的「道德目的」作為可信賴（trustworthy）人工智能的組成要素之首，使人工智能為個人和社會的福祉而發展。但這種「道德目的」不應該僅表現為單向的人工智能對人類權利的尊重與規範的遵守，還應同時考慮人類對智能機器人相應的保護，將智能機器人的主體性考慮在內，要求自然人同樣尊重智能機器人的某些重要利益或重要活動，減少人類對智能機器人的非人道的、無序的行為。從刑法的角度來看，未來刑法不僅會規制智能機器人的犯罪行為，也會打擊針對智能機器人的犯罪行為。

受侵害人工智能體的刑法保護與救濟路徑

現階段：借用行政犯的立法模式來保護人工智能體

當前，法定犯（statutory offence）時代的到來可謂是一種社會發展的必然趨勢。應該說，法定犯的規定，並非像自然犯一樣具有高度的民眾認同度，大多數情況下，其是基於一定時期內維護社會秩序的需要而被立法者制定的。換言之，法定犯不是為了直接地保護某個具體受害人的個別利益免受損害，因為法定犯所規制的行為本身並不具有強烈的道德可譴責性，典型的法定犯立法模式往往規定了「違反……某行政法規的規定」之類的空白罪狀，從而與行政法規產生了內在聯繫，應該說，不論是空白罪狀中的行政法規還是法定犯本身，最終都是為了國家的管理制度能夠有序運行而存在，其背後所維護的是超個體的集體法益。換言之，法定犯所保護的法益並非具象的人或物的狀態，而是該狀態背後的法秩序。

就智能機器人受侵害的問題而言，短時間內要讓智能機器人獲得刑法上的主體資格並不現實，但是，這並不影響刑法規制對智能機器人的侵害行為，原因在於智能機器人所嵌入的社會秩序與價值體系本身已經達到了刑法介入的程度，事實上也早已處於刑法所保護的範圍。具言之，如前所述，智能機器人將承載人類愈來愈多的重要道德情感和重大利益，對智能機器人的某些破壞或干擾行為，將可能直接導致人類的這些重要利益受到損失，即使還沒有獲得法律主體資格的承認，刑法也可以將這部分侵害行為，以典型法定犯的形式專門規定下來，以避免對人類利益保護的不周延。

或許有人認為當前的罪名可以涵蓋一部分通過侵害智能

機器人進而危害人類的行為，如故意破壞自動駕駛汽車的控制系統，使其在行駛過程中足以發生傾覆、毀壞的危險，可以認定行為人觸犯了破壞交通工具罪，但是如前所述，隨着自動駕駛技術智能化和自動化的提升，未來自動駕駛汽車不會被簡單定義為交通工具，而是通過獲得極大自主性而具備主體資格的自動駕駛機器人，因為其在極大程度上脫離程序控制獨立駕駛發生交通事故後，完全可能由其獨自承擔法律責任，這將與傳統意義上的以客體物為存在形式的「交通工具」產生本質性差異，即只有那些非智能或弱智能的火車、汽車、電車、船隻和航空器，才會被解釋為破壞交通工具罪中的「交通工具」。

此外，諸如當前已經出現的利用性愛機器人開設妓院的行為是否合法？若不合法，是否應當受到刑法追究？是否可以適用已有的組織賣淫罪、聚眾淫亂罪等罪名對此行為進行規制？此等問題均存在較大爭議。這說明，智能機器人的出現對傳統刑法的理論與實務提出了較大挑戰，刑法當前的罪名無法始終或準確涵蓋針對智能機器人的侵害行為，忽視智能機器人技術的進步而堅守舊的規則，難以適應社會的新發展和新要求。

如果說，刑法以主動的姿態介入規制嚴重侵害智能機器人的行為是基於一種人本主義的功利性目的的話，則應首選法定犯的規定模式，因為一方面法定犯不過分關注其所直接保護的具象的人或物的狀態的本然性質，從而有利於迴避「智能機器人是否應具有主體資格」的爭議；另一方面，就立法技巧而言，其規制模式具有便利性和靈活性，因而使得此類犯罪的規定更具有包容性，從而令刑法能夠始終保持與科技發展的及時雙向互動。最後，從反面看，如果法定犯的立法模式成立的話，勢必意味着存在行政法的前置性保護，

這將使智能機器人及其所承載的重要人類利益受到雙重保障。

綜上所述，本文認為，可以先制定一部類似於《智能機器人管理與保護條例》的行政法規，以明確規定享有法律保護的智能機器人的定義，如「本條例所稱『機器人』是指以服務社會為目的，能夠獨立行為，具備良好的類人性的感知和控制系統，可以通過大數據處理進行深度學習和分析的類人型智能機器人。」並規定不得出現針對智能機器人的各項不良行為，例如「不得隨意肢解、損毀智能機器人；不得非法干擾智能機器人正常工作；不得利用智能機器人做違背公序良俗的事情……」等，並以罰款、拘留等形式確保該條例得到貫徹落實。進而《中華人民共和國刑法》可以在總則第五章「其他規定」中，增加刑法中智能機器人的定義，具體可以參照《智能機器人管理與保護條例》的行政法規中對智能機器人的定義內容；在分則中設立專章，規定有關智能機器人受侵害的犯罪，可以在具體條文中，均以「違反保護智能機器人的規定」為前提條件，輔之以行為要件和情節要件，作為該條文的要件內容，例如，「違反保護智能機器人的規定，公然與智能機器人進行性交，情節嚴重的，處……；情節特別嚴重的，處……。」當然，就這些新罪名的法定刑配置而言，究竟有沒有必要參考針對人類類似侵害行為的刑罰類型和幅度，還是無需參考，而徑直設置這一類行為的刑罰域，還有待進一步商榷。

未來展望：直面人工智能體自身法益的保護模式

如果說上述現實保護模式是以保護人工智能體所嵌入的社會秩序為基底的間接模式的話，那麼，在未來，智能機器人的主體地位獲得法律的認同之後，刑法可以拋開前置性的

行政規範，徑直對人工智能體展開保護。具言之，本文認為，智能機器人理應獲得法律上的主體身分認同，其自身的利益訴求，也會隨着法律主體資格的確認而被刑法所重視，成為獨立於人類的利益。姑且將智能機器人的那部分自身利益，即那些類似於人的生存權一樣，能夠維持智能機器人持續地存在、運作（以後或許稱為工作）的利益，包括前文論述的不受無端破壞、肢解等，有獨立作出判斷、選擇和創作的自由，以及收集、保有重要數據等利益，稱為智能機器人的核心利益。本文認為，智能機器人保護的非人本主義利益觀，要求我們將智能機器人的核心利益納入道德和價值範疇，而不再將破壞智能機器人核心利益的行為視為破壞人類利益之附屬的行為，即應該對這類危害性的行為予以獨立的價值評價，使其成為刑法上可以脫離於人類利益而評價的對象。

Raffaele Garofalo 將典型的犯罪（自然犯）認定為是對人類憐憫情操和正直情操的違反（參見加羅法洛 , 1996:44），是一種顯而易見的罪惡，如果將智能機器人的重要利益評價為獨立於人類的重要利益，那顯然可以將損害這部分利益的行為視為一種顯而易見的罪惡。也即智能機器人獲得法律主體地位的承認後，刑法的人本主義利益觀才會徹底改變，智能機器人才會擺脫人類的附屬物性質，真正以犯罪者和犯罪受害者的角色登上刑法舞臺。此時，保護智能機器人切身利益的驅動力，從原來的以人類利益為中心的功利性需求，轉變為發自內在的道德請求，那些掙脫道德約束而嚴重侵害智能機器人切身利益的行為，因為徹底違背主流道德觀而受到刑法的譴責，此時，如肢解智能機器人並使之完全喪失行為能力的行為，並非因為使人類失去了重要的勞動工具而被規定為類似財產性質的犯罪行為，而是出於該行為與殘忍地殺害人類的行為無異，是具有人身損害性的犯罪行為。

對此，刑法應重新調整侵害智能機器人的犯罪規

定——在前述法定犯模式的基礎上，將侵害獲得主體資格的智能機器人之核心利益的行為獨立出來，並在《中華人民共和國刑法》總則第五章關於智能機器人的定義後面，增加其獲得法律主體資格的標準，即在民政部門或科技部門獲得類公民主體身分而有備案登記的類人型智能機器人。被單獨分離出來的這部分犯罪行為根據法益類型，可以在《中華人民共和國刑法》分則有關智能機器人犯罪專章中，具體分為破壞智能機器人完整性的犯罪、干擾智能機器人行動自由的犯罪，和嚴重阻撓智能機器人收集、保留重要數據，修改、竊取其重要數據犯罪等三種類罪名。同時，對於那些沒有直接危害到智能機器人核心利益的其他侵害行為，如剽竊創作型機器人的音樂作品，並以此獲取數額較大的收益，由於直接威脅或損害到的是著作權秩序下的人類的利益，刑法仍然應該以法定犯的形式予以規定。

當然，隨着智能機器人技術的發展，智能機器人的核心利益類型還會不斷擴張，比如未來在智能機器人與人類能夠友好共處的前提下，如果最終允許人類與智能機器人登記「結婚」的話，那麼，也可能會產生智能機器人的婚姻自由；如果智能機器人有相互交往的需求的話，那麼，勢必會產生通信自由；如果允許智能機器人參與人類社會的管理或智能機器人群體內部的管理的話，那麼，很可能會延伸出智能機器人的政治選舉自由、信仰自由等。此時，對智能機器人刑法保護的正當性依據，可能更多地是取決於對行為本身的客觀評判，而非立法者基於人本主義的考量（白建軍，2018）。

本文認為，這時刑法勢必也會相應地逐漸增加新的罪名規定，保護智能機器人的罪名類型，將會出現藉由保護社會秩序來保護智能機器人的間接保護模式，向直接保護模式拓

展的趨勢。此外，侵害智能機器人核心利益的犯罪行為，還因為智能機器人嵌入社會秩序和價值體系後，與人類的利益相互牽連，因而表現出複雜的社會危害性，刑法對於這部分犯罪行為的打擊態勢也會更為嚴厲。

對智能機器人遭遇刑事被害的補償與救濟

法諺云：無救濟，則無權利。對未來獲得主體地位智能機器人的利益保護，應該比照對人類權利的保護，除了應該盡可能將智能機器人的利益納入刑法和其他法律範疇予以事前保護外，也應該考慮對那些受犯罪侵害的智能機器人的利益進行事後救濟。Christopher Stone 提出，某一主體能否擁有法律權利應滿足以下條件：第一，該主體應其要求可以提起法律訴訟；第二，法院在決定授予法律救濟時必須考慮到損害；第三，法律救濟必須滿足它的利益需求（McNally and Inayatullah, 1988:126）。

在恢復性司法（restorative justice）理念下，現有的刑事犯罪所要考慮的救濟問題，着重於對被害人的賠償問題，並認為刑事損害賠償不僅能使被害人的利益獲得實質性的保護，對預防犯罪也有一定作用。[7] 刑事賠償包括物質賠償和非物質賠償，當前關於被害人賠償的主體包括犯罪人賠償和國家賠償。在中國，針對前者，主要通過被害人或其法定代理人、近親屬提起附帶民事訴訟的方式，要求犯罪人賠償，

7. 美國全國少年司法研究中心（National Center for Juvenile Justice）曾在猶他州調查了6,336件官方統計的少年假釋案件。結果發現，賠償的使用與一些少年犯罪人中累犯的行為的顯著減少有正面聯繫（U.S. Department of Justice, 1992:4; 李偉, 2014:208）。

而智能機器人脱離於自然人屬性，加上侵害行為往往可能嚴重損壞智能機器人的智能感知系統，所以其獨立提出附帶民事訴訟的能力不足，對此，檢方或與智能機器人關係密切的其他公民，可以根據智能機器人的修復情況，代替其決定是否提出附帶民事訴訟請求，具體的賠償金額，可以要求犯罪人返還對智能機器人剝奪的經濟利益，或者參考修復被損智能機器人的費用，以及日後加大保養的費用，由智能機器人的所有者或管理者代為保管使用；亦或要求犯罪人在限期內重新修復智能機器人、給智能機器人賠禮道歉等。此外，國家也應該就某些不當的國家行為，對智能機器人造成的重大利益損害進行賠償，尤其是涉及到對智能機器人本身與財產性的利益損害時，應當參照自然人賠償標準，對智能機器人進行彌補。

餘論與展望

本文所討論的主旨問題，是刑法究竟應該如何面對和保護人工智能體，本文的基本觀點是人工智能體最終應該具有獨立的被害人地位。但其實真正讓人工智能體能夠獲得法律主體性承認的，一方面是人工智能體能否在未來社會中發揮愈來愈重要的正面積極作用；另一方面，更深層次的是人類能否突破人類中心主義思想，對「自我」與「他者」之間的關係進行更為深刻的反思。

應該說，現有的刑法理論是在人本主義層面上建立起來的。人類經過漫長的探索，才擺脱神靈的控制，進而認識到人自身的意義與價值，人是自由的，這意味着人是自己的主宰，社會定紛止爭的權力應該是保護人，並由人來制定與執

行的，而刑法是執行人類意志最強有力的規範武器，所以刑法一開始就是由人所制定，並為人類服務的法律。人工智能體作為人類的生成物，是與人類不同的「他者」，是現有刑法制定主體與調整對象之外的存在物，站在人類中心主義立場看，除非人工智能體被視為與人類財產一樣的附屬物，否則其幾乎無法與刑法有任何實質的交集。然而事實並非如此，人類中心主義思想已經在日益嚴重的環境污染和生態破壞的現實情形下暴露出了自身的弊端與局限性，當下人工智能體的社會屬性早已超越了其自然屬性，從關係本體論出發，人類的整體幸福感未必不可以建立在與自身以外之存在物的廣泛交互性上，這必然促使人類更加注重「他者」對於社會價值與功能的發揮。刑法也應該維護人工智能體這種「他者」的社會價值，如同環境對人類的價值需要，由刑法對那些破壞生態平衡的嚴重危害行為施以懲罰來保護一樣。

當前刑法面臨抉擇之處就在於能否以及如何突破人類中心主義的立場，給人工智能體這樣的非人類存在物之「他者」一個庇護之所。或許古典功利主義理論（utilitarianism）能提供一種現實的解釋途徑。因為功利主義理論仍然以人類中心主義為基礎的，只要能夠「實現最大多數人的最大幸福」，就是值得肯定的，以此為目標建立的制度就是合適的制度，保護人工智能體可以被解釋為是為了維護人類的最大利益，可以提升人類的整體福利，所以刑法保護人工智能體是合適的。這樣刑法就可以在保有原來立場的同時，也能給予人工智能體一定程度的制度保護。但沒有突破人類中心主義立場，就意味着對人工智能體的保護受限於人類的需求與利益，只要不涉及人類的利益，人工智能體就不在刑法的保護範圍內，這樣與其說是保護人工智能體，不如說是在保護人類自身，進而「自我」與「他者」之間的矛盾仍然沒有得到調解。

本文最後展望，未來的刑法會突破人本主義的制約，將人工智能體這樣的「他者」當作人類自身一樣的存在物去對待，這既是人類自我觀念轉變的結果，也是人類文明進一步開化的結果。屆時刑法將既保護人類自身，又平等地保護諸如環境、動物和人工智能體一樣的非人類存在物。

參考書目

A. R. 拉德克利夫 • 布朗。2014。《原始社會結構與功能》，丁國勇譯。南昌：江西教育出版社。

中國日報網。2016。〈全球金融業進入『機器人時代』：你的血汗錢是否安全？〉，3 月 21 日。取自：http://m.cankaoxiaoxi.com/finance/20160321/1105779.shtml。

王肅之。2018。〈人工智能犯罪的理論與立法問題初探〉，《大連理工大學學報（社會科學版）》，第 39 卷，第 4 期，頁 53–63。

王耀彬。2019。〈類人型人工智能實體的刑事責任主體資格審視〉，《西安交通大學學報（社會科學版）》，第 39 卷，第 1 期，頁 138–44。

加羅法洛。1996。《犯罪學》，耿偉、王新譯。北京：中國大百科全書出版社。

司曉、曹建峰。2017。〈論人工智能的民事責任：以自動駕駛汽車和智能機器人為切入點〉，《法律科學（西北政法大學學報）》，第 5 期，頁 166–73。

白建軍。2018。〈法定犯正當性研究：從自然犯與法定犯比較的角度展開〉，《政治與法律》，第 6 期，頁 2–12。

皮勇。2018。〈人工智能刑事法治的基本問題〉，《比較法研究》，第 5 期，頁 149–66。

托比•沃爾什。2018。《人工智能會取代人類嗎？》，閭佳譯。北京：北京聯合出版公司。

江山。2000。〈法律革命：從傳統到超現代——兼談環境資源法的法理問題〉，《比較法研究》，第 1 期，頁 1–37。

李偉編。2014。《犯罪被害人學教程》。北京：北京大學出版社。

杜嚴勇。2014。〈現代軍用機器人的倫理困境〉，《倫理學研究》，第 5 期，頁 98–102。

阿西莫夫。2005。《機器人短篇全集》，漢聲雜誌譯。北京：天地出版社。

時延安。2018。〈犯罪化與懲罰體系的完善〉，《中國社會科學》，第 10 期，頁 102–25，206–07。

高奇琦。2018。《人工智能：馴服賽維坦》。上海：上海交通大學出版社。

張玉潔。2017。〈論人工智能時代的機器人權利及其風險規制〉，《東方法學》，第 6 期，頁 56–66。

張愛萍。2016。〈與機器人談『感情』，人類是否『很受傷』？——也談人工智能與人類情感融合的前景〉，光明網，4 月 5 日。取自：http://m.cankaoxiaoxi.com/science/20160405/1118962.shtml。

曹明德。2002。〈法律生態化趨勢初探〉，《現代法學》，第 24 卷，第 2 期，頁 114–23。

曹明德。2007。《生態法新探》，第 2 版。北京：人民出版社。

梁根林。2005。《刑事法網：擴張與限縮》。北京：法律出版社。

麥克•馬圭爾、羅德•摩根、羅伯特•賴納等。2012。《牛

津犯罪學指南》，劉仁文等譯。北京：中國人民公安大學出版社。

焦艷鵬。2012。《刑法生態法益論》。北京：中國政法大學出版社。

搜狐。2016。〈《西部世界》這部 9.2 分的『小黃片』憑什麼封神〉，11 月 3 日。取自：https://www.sohu.com/a/118039969_520625。

搜狐。2018。〈剛剛，歐盟 AI 道德準則草案出爐！可信賴的 AI 才能成為人類的北極星〉，12 月 19 日。取自：https://www.sohu.com/a/282938991_354973。

瑞恩 • 卡洛、邁克爾 • 弗魯姆金、伊恩 • 克爾編。2018。《人工智能與法律的對話》，陳吉棟、董惠敏、杭穎穎譯。上海：上海人民出版社。

劉憲權編。2018。《人工智能：刑法的時代挑戰》。上海：上海人民出版社。

鄭玉雙。2016。〈為犯罪化尋找道德根基：評范伯格的《刑法的道德界限》〉，《政法論壇》，第 34 卷，第 2 期，頁 183–91。

Gleß, Sabine and Thomas Weigend. 2014. “Intelligente Agenten und das Strafrecht” (Intelligent Agents and Criminal Law), *Zeitschrift für die gesamte Strafrechtswissenschaft*, 126(3):561–91.

McNally, Phil and Sohail Inayatullah. 1988. “The Rights of Robots: Technology, Culture and Law in the 21st Century,” *Futures*, 20(2):119–36.

Palila v. Hawaii Department of Land and Natural Resources. Retrieved from: https://elr.info/sites/default/files/litigation/17.20514.htm.

Scheutz, Matthias. 2012. “The Inherent Dangers of Unidirectional Emotional Bonds between Humans and Social Robots,” in

Patrick Lin, Keith Abney and George A. Bekey (eds.), *Robot Ethics: The Ethical and Social Implications of Robotics*. Cambridge, MA: The MIT Press, pp. 205–21.

U.S. Department of Justice. 1992. "Resititution and Juvenile Recidivism," *Juvenile Justice Bulletin*, September. Retrieved from: https://www.ncjrs.gov/pdffiles1/Digitization/137774NCJRS.pdf.

被害者視角

刑法何以保護人工智能體？

摘要

人工智能體在廣泛地參與人類社會生活過程中，其獨特的利益訴求、與生俱來的利他性，以及未來獨立承擔法律責任的可能性，使得其理應獲得類似於人類主體的法律主體地位。人工智能體將承載愈來愈多的人類道德情感，並與社會秩序產生日益緊密的聯繫，這些特徵強化了其與刑法的內在聯繫。刑法應在主體性視角下保護人工智能體自身及其承載的重要利益，應將對人工智能體的某些非道德或無序行為納入犯罪範疇。當前可先以典型法定犯的模式規定此類犯罪，待未來人工智能體的法律主體地位得到普遍承認時，再逐漸向直接保護的模式轉型，並應在恢復性司法理念支配下，對受侵害的人工智能體予以必要救濟和補償。

The Victim's Perspective
How Does Criminal Law Protect AI?

Jia Jian

Abstract

Artificially intelligent entities are widely involved in the life of human societies. Their unique interests and demands, their innate altruism, and the possibility that they might in the future be able to assume independent legal responsibility will cause them to attain a legal status similar to that of humans. More and more, artificially intelligent entities will embody human morals and emotions and be increasingly closely linked with social order—characteristics that will strengthen the inherent connection with criminal law. From the perspective of subjectivity, criminal laws should protect artificially intelligent entities themselves and the important human interests that they embody. Certain unethical or disorderly behaviours towards artificially intelligent entities should be criminalized. For now, such crimes can be treated as typical statutory offences. In the future, when artificially intelligent entities are generally recognized as legal subjects, a gradual transition can be made to a direct protection model. Under the concept of restorative justice, the artificially intelligent entity that has been the victim of a crime should be given necessary relief and compensation.

Hong Kong Institute of Asia-Pacific Studies

The Hong Kong Institute of Asia-Pacific Studies (HKIAPS) was established in September 1990 to promote multidisciplinary social science research on social, political and economic development. Research emphasis is placed on the role of Hong Kong in the Asia-Pacific region and the reciprocal effects of the development of Hong Kong and the Asia-Pacific region.

Director:

Fung, Anthony Ying-him, PhD (Minnesota),
Professor, School of Journalism and Communication

Associate Directors:

Hong, Ying-yi, PhD (Columbia),
Choh-Ming Li Professor of Marketing

Ng, Mee-kam, PhD (UCLA),
Professor, Department of Geography and Resource Management

Zheng, Victor Wan-tai, PhD (University of Hong Kong),
Associate Director (Executive), HKIAPS

ISBN 978-962-441-244-4