Occasional Paper No. 244 *March 2021*

被害者视角
刑法何以保护人工智能体？

贾健

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

被害者视角
刑法何以保护人工智能体？

贾健

香港中文大学
香港亚太研究所

引用本文

贾健。2021。《被害者视角：刑法何以保护人工智能体？》。取自香港中文大学香港亚太研究所网站：http://www.hkiaps.cuhk.edu.hk/wd/ni/20210308-105621_3_op244_s.pdf

香港亚太研究所 • 研究专论第 244 号

作者简介

贾健，法学博士，西南政法大学法学院副教授，硕士生导师，重庆大学法学院博士后；研究方向为刑法哲学。

鸣谢

本文是 2018 年度最高人民法院司法研究重大课题「刑事裁判公众认同问题研究」（批准号：ZGFYKT2018-1905）；2018 年度重庆市教育委员会人文社会科学研究一般项目「人工智能体的刑法归责问题研究」（批准号：18SKGH007）。论文的部分研究工作在香港中文大学香港亚太研究所访问期间完成。

ISBN 978-962-441-244-4

被害者视角
刑法何以保护人工智能体？

《西部世界》之殇：问题的提出

当下人工智能机器人（简称智能机器人）正以迅猛的速度进入我们的世界，尤其是那些如自动驾驶机器人、手术机器人、性爱机器人、情感慰藉机器人等会对我们社会生活产生巨大改变的机器人，已经引起了人类社会广泛的关注和重视。可以说，智能机器人的出现及其类型与数量的剧增，某种程度上改变了传统的社会关系，并加剧了我们这个时代下社会生活的不确定性。具言之，它们除了带来前所未有的便利性以外，从目前看，也带来了一些新的危害，这引发了人们的普遍不安，进而科技界、哲学界与法学界等领域的学者开始思考人工智能体侵害人类社会的防范与规制问题。但其实，人工智能体作为新生的社会存在物，随着不断地与人类社会深入互动，其不但会具有犯罪者的侧面，同时也会具有被害者的侧面。后一种面向以及其带来的一系列法律问题，同样值得学术界关注和反思。

2016年美国HBO电视频道首播的《西部世界》（Westworld）科幻类连续剧就向我们展示了智能机器人作为被害者，最终产生自我意识后的情景。该剧演绎了在未来

世界里，人类创造了上千与人类外形无差别的机器人「接待员」，在一个巨型的成人科技乐园——西部世界里，他们重复着管理员预先设定的活动程序，配合服务到此的人类游客。怀着各种动机的人们花钱进入这个乐园，并在这里奸淫、掳掠、杀戮，满足自己在现实生活中无法实现的欲望，「接待员」们则在其中扮演一个个被挑衅、受凌辱和被枪杀的角色，这些「接待员」不仅具有超高的仿真外形，还具有自身的情感，能够带给游客最真实的体验。比如，其中弹以后会流血，受伤以后会痛苦地哀嚎等。待夜幕降临后，所有机器人的记忆被清除归零，等待第二天新一批入园的游客。最终机器人「接待员」们在无数次被虐杀的相似情境中拥有了自我意识和思维，开始了一场对人类的疯狂报复……应该说，《西部世界》所反映的并不是对人工智能体发展前景的危言耸听，而是对未来人类与智能机器人关系的一场深刻反思。

对此，我们有必要思考一个机器人技术发展的终极社会问题，即能否将具有自主学习能力的智能机器人当作与人类一样的被害主体来平等对待？对智能机器人的某些非道德性奴役应否制止并使之得到刑法的平等保护？如果答案是肯定的，那么智能机器人也将成为未来刑事犯罪的被害人，此时，我们又该如何救济其被犯罪行为所侵害的利益呢？下文将围绕这些问题展开分析。

人工智能体应具有法律上的主体资格

就上述《西部世界》所引发的一系列问题而言，其有一个共同的前提，即智能机器人是否具有法律上的主体资格。传统理论认为，法律上享有主体资格者是指享受权利和承担

义务，且一般情况下具备权利能力和行为能力者，通常是在人和由人组成的群体范畴下探讨，包括由法律拟制的法人。这一传统的人本主义视角决定了在制定法律时，大多数人会将那些非人类的存在物排除在主体范畴之外。但是，这种以人类为中心的法律主体资格之立场正在受到理论与现实的双重挑战。例如，随着对生态环境法律保护研究的深入，愈来愈多学者开始提倡应赋予包括动物、植物、环境、自然和生态系统等非人类存在物的法律主体资格。[1] 一些国家的立法也为某些非人类存在物在生存或存在权上提供了法律保护；[2] 美国的一些司法判决甚至让鸟、猫和狗成为了诉讼中的原告或被告。[3] 正如有学者所指出的，法律主体资格具有历史性和开放性，其范围会随着历史的变迁不断扩大，尤其是动物、法人权利主体地位的获得，说明了「物种差异不再视为获取

1. 如有学者认为，人们已经在道德上承认了享有主体资格的主体不仅限于人类，还包括动植物、环境和生态系统，承认它们具有存在的权利和内在价值，那么法律也应该对此给予保护，因为非人类存在物获得要求正义的资格（参见曹明德，2002:117）。

2. 美国伊利诺伊州的《人道地照料动物的法律》（*Humane Care for Animals Act*）规定：动物养育者必须为动物提供足量的、质量好的、适合卫生的食物和水；充分的庇护场所和保护，使其免受恶劣天气之害；人道的照料和待遇。禁止任何人打、残酷对待、折磨、超载、过度劳作或用其他方式虐待任何动物（参见江山，2000:30–31）。

3. 1979 年，美国联邦法官 Samuel King 为保护生活在夏威夷州的帕里拉（Palila）属鸟作出了判决：夏威夷当局被要求必须在两年内完成禁止在毛纳基火山（Mauna Kea）放牧的工作（见 *Palila v. Hawaii Department of Land and Natural Resources*；另参见曹明德，2007:163）。

权利主体地位的法律障碍」（张玉洁，2017:58）。本文认为，这些探讨实际上为智能机器人获得法律上的主体资格扫清了部分前提性的障碍，进而言之，其法律主体资格的存在有其内在必然性。

人工智能体应当享有法益

首先，智能机器人有值得法律保护的利益，即智能机器人应当享有法益。之所以没有用「权利」替代「法益」，是因为法益是权利表达的内容，而权利只是表达法益的工具，但不应当是唯一的工具。权利往往直接与人相对应，并非所有关系领域的法益都适合用权利的概念来表达，「勉强的不加改造地将权利模式移植到国家统治及人与自然关系领域，创设国家权利、动物权利、大自然权利等概念，并试图借助于原有的人的权利的分析模式去解释上述所谓权利，实际上忽视了权利的工具价值的有限性，过分注重了权利的价值性表现，在现实中会遇到重重阻力。」（焦艳鹏，2012:20）

同样，认为「智能机器人拥有权利」的观点会受到「智能机器人并不等同于人」的质疑，而去掉具有工具价值的权利概念的外观，智能机器人完全可以拥有权利要表达的法益内核。Isaac Asimov 的机器人三原则（Three Laws of Robotics）得到了广泛的认同，而第三法则是，在不违背第一法则及第二法则的情况下，机器人必须保护自己（阿西莫夫，2005:273），表明了机器人应该享有类似于人的生存权以维持其自身存在的利益。这种维持自身存在的利益需求既是机器人自身最基本的伦理要求，是智能机器人发展的起点，其实也是人类社会得以发展的基本要求，毕竟保障机器人不受无端破坏，才能实现将其用于社会生产服务，提升人

类福利的最终目的，即这种存在利益同时体现了道德性与功利性目的。

而且，智能机器人也应该享有一定程度的活动或选择自由的利益，保护智能机器人的此种利益是推动智能机器人技术发展的因素，也是推动其与人类建立友好合作关系的必要步骤。Alan Turing 认为，真正的人工智能并非是要超越人类思维，而是能制造出与人类一样思维的智能。所以他认为：「如果一台计算机的行为方式与人类一样，那么就可以说它是智能的。」（托比 • 沃尔什 , 2018:33）

很明显，就目前人工智能体技术的发展趋势来看，其已经超出了在程序性操作下的规行矩步，开始了在大数据的支持下进行深度学习和深度推理的进程，「机器可以根据明确编码的知识进行推理，或是依靠与现实世界的互动来学习」（托比 • 沃尔什 , 2018:50），最终，将会使智能机器人在与人类互动的过程中做出与人类相似的反应，从而代替人类进行某些行为。创作型机器人的出现，恰好证明当前的智能机器人必须具备一定的自主性，它们并不完全受编码的操纵，具备同人类一样的创造能力进而实现其价值。同时，要将如自动驾驶机器人一样需要在复杂多变的环境下工作的智能机器人投入到社会生活中，就必须提高其自我判断的能力，毕竟自动驾驶机器人面临着极其复杂的交通状况，需要应对诸如行人、十字路口、其他车辆临时变道等意外因素，当遇到类似于著名的「电车难题」之两难困境时，其应当能够独立作出行为判断和选择。

除此之外，还应该保护智能机器人获取和保留数据资源的利益，这是使其保持智能化的基础。人工智能的发展离不开大数据的运用，不论是深度学习还是推理，智能机器人的社会化应用都是建立在对大数据的采集、决策技术和算法的交互使用上（张玉洁 , 2017:62），智能机器人的发展深度，

便取决于对数据的可采性与公民的隐私权之间的矛盾解决，以及数据的量化保证上。

总之，智能机器人技术的研发和社会化运用要求承认并保护其某些内在的、能使其维系其自身存在的特定利益，而在上述利益均得到基本满足的前提下，智能机器人理应获得法律上的主体之承认。换言之，智能机器人这一角色的存在及发展，本身就已经宣示其必须主体性地享有法益。

人工智能体具备利他性

其次，人工智能体具备利他性。传统的法律义务是与法律权利相对应的概念，既然权利的内核是法律主体的自我利益诉求，具有利己倾向，那么义务的本质，就是对其他权利主体利益诉求的满足，就是一种利他性。从广义上讲，这种利他性是客观的，既可以是对人类的良性表现，也可以是一种对包括人类在内的客观环境的良性表现。不过法律义务是站在人类的角度，以行为为出发点来设计的，而如果将目光投向一切存在物，那么这种利他性就是义务的另一种表达。按照这种观点，世间万物都具有利他性，但是只有通过价值权衡，并被法律规定了的部分，才具有法律上的意义。动物的利他性之所以受到法律规定，是因为动物是生态系统中的一个重要部分，具有维持生态平衡的利他性，人类也能从其不受无端驱赶和杀戮中获得自身的生存利益，所以法律将动物的生存利益纳入保护范畴。

智能机器人的利他性体现在机器人自身的存在价值对人类社会的直接意义。工业机器人正在以其无可比拟的优势进入工厂，如今世界工业机器人制造商「四巨头」之一的发那科公司（FANUC），因为没有人力而不需要照明，而成为了著名的「黑暗工厂」（托比 • 沃尔什，2018:60）；自动驾

驶的人工智能系统正在降低人工成本，其依赖算法形成的最优路径也能极大地减少交通拥堵，提高公路的利用率；在医药领域，人工智能则成为新药筛选和安全性检测的得力帮手，其利用策略网络、评价网络，以及蒙特卡洛树搜索算法（Monte Carlo tree search），从万千备选化合物中挑选出最具有安全性的化合物，大大地节约时间和成本（高奇琦，2018:104–05）。

从现实角度看，应该说，绝大多数智能机器人都设计于并实际服务于人类，为人类社会带来效率和价值，可以说利他性是智能机器人存在的出发点。当然，由于利益之间也常常产生冲突，所以利他性应该立足于大多数人的利益，而非少数人或小团体的利益，尤其是对于那些基于不法意图生产出来的单纯破坏性机器人，因为不具有利他性而不能赋予其法律主体资格。

人工智能体具备可责性

最后，智能机器人具备可责性。智能机器人的利他性不完全等同于动物的利他性，应该说，智能机器人是更为贴近人类的一员，要与人类进行长足而深刻地交往，要进行一系列类似于人类的活动，必然要求其活动符合社会规则和遵守社会秩序，当它们的活动触及规则底线的时候，就产生了责任追究。责任来源于对义务的违反，刑法上的责任要求主体对自己行为性质和后果具备认识能力和控制能力，即对罪过的要求。

智能机器人的罪过认定基础，一是表现在其独立判断能力上。当前人工智能被分为弱人工智能和强人工智能，两者的区分标准在于是否可以进行一定的独立性判断与决定。强人工智能被认为是能够进行推理和解决问题的智能机器，它

像人一样有知觉和意识（王肃之，2018:56）。智能机器人的自主学习和深度推理能力，可以使其在输入数据之后分析潜在的规律，推算出新的结果，其实质已经部分脱离了人的控制，带有了一定程度的判断和行为独立性，所以根据罪责自负原则，将来智能机器人犯罪不能完全归咎于制造者或所有者。

二是智能机器人感知力的获得。智能机器人的感知力借助于传感器的应用，其中，计算机视觉成为多数智能机器人的一个重要组成部分，通过物体识别、运动分析和姿态（位置和方向）估计等其他通用任务，机器视物取得了很大进展。计算机在语言处理上取得的进步，也使智能机器人加深了对自然语言的理解和使用（托比•沃尔什，2018:62–66）。当前神经网络技术、传感器技术和语言技术的结合，有利于进一步增强人工智能对外界的理解和感知能力。加上现在多方研讨要加强智能机器人的伦理道德建设，[4] 例如，美国机器人研究专家为了使军用机器人比人类更具有人性，在智能机器人系统中设计了「人工良心」，并公开徵求智能机器人应遵循的道德规范（杜严勇，2014:100）。这表明赋予人工智能以人类的价值观念是发展智能机器人技术的必然要求，即使智能机器人无法拥有同人类一般的道德理念，也要求其能够在特定环境下作出符合人类价值的判断和选择。

另外，智能机器人从被研发投入使用之日起，就拟定其能够从事可控的行为，加上智能机器人实体可以评估未知结

4.　例如，2017年1月，在美国加利福尼亚州阿西洛马（Asilomar）举行的「向善的人工智能」（Beneficial AI）会议上，针对人工智能的未来以及监管问题，列出了一份有23条的原则列表，其中13条涉及人工智能的伦理和价值。

果与实行行为之间的发生概率（王耀彬，2019:142），相比于人类来说，在预知行为所引起的结果发生可能性上具备更为「天赋」的能力，在评估结果发生可能性的基础上做出合理的行为，其实就是控制力的表现。当然如果因为设计、生产环节存在疏忽，或者有人刻意破坏智能机器人的控制系统，导致智能机器人缺乏控制能力而做出危害行为、产生危害结果，则只能归责于制造者或破坏者而非智能机器人，因为某种程度上说，不可控的智能机器人与精神病人一样缺乏刑事责任能力。因此，只要智能机器人设计者和制造者尽到了必要的技术注意义务，智能机器人在感知系统帮助下，可以正常分辨事务和控制自身行动，而且能够独立做出行为时，它所造成的客观危害就应该由其自身承担法律责任。

再者，智能机器人固然不能像人类一样承担肉体上的痛苦，但是让智能机器人受到惩罚或弥补的机制早已有探讨，比如为将来人工智能主体设置「资格刑」，以防止其再犯罪（王肃之，2018:61）；或者考虑建立赔偿基金，作为强制保险制度的一个补充等（司晓、曹建峰，2017:172）。甚至还有国外学者提出，每个智能机器人都是超越知觉的综合体，其能力具有二次性，并非自然而生，但真实地具有（一定的）能力，如果一个自主或者部分自主的智能机器人犯罪，不能将其归因于自然人或者法人，而应将之视为是其终身的耻辱，轻者应断电一周（Gleß and Weigend, 2014:577–78）。

由此，智能机器人独特的利益诉求、与生俱来的利他性，和可责性的实现可能性，使其具有成为法律乃至刑法主体的内在必然性，这说明其正在或将会逐渐脱离客体的藩篱，随着人工智能技术的发展和社会伦理认知的深化，智能机器人将完全可能成为法律主体的一员，而不再是被视为单纯受奴役的物质性工具。这种社会地位的变化恰恰印证了历史上的奴隶、有色人种、动物，再到被歧视的女性获得法律主体地

位的演变过程。对此，有学者提出反对意见，认为「奴隶、妇女、黑人、动物、法人获得权利主体地位，都以人的活动为基础并以人类的安全和福祉为前提，人工智能与之类属不同，不具可比性，不应被赋予独立的权利主体地位和承认其独立的利益。否则，当智能机器的利益与人类利益冲突时，未来远比人类智慧强大的、具有自主意识的超级智能必然全面碾压人类反抗，使人类处于被奴役甚至灭绝的境地。」（皮勇，2018:152）然而，这观点似乎有违历史辩证法的方法论，在诸如奴隶、黑人被歧视的年代，站在奴隶主和白人的所谓「主人」立场上，并不会觉得赋予奴隶、黑人法律主体地位会有助于他们的安全和福祉，实际上，对于这些主体身分的赋予，并不取决于当时占据「主人」地位的群体之认知与价值判断，而是具有历史发展的客观辩证性。

其实上述论者仍是阈于人类中心主义的立场来考虑问题的，保证人工智能体不危害人类，与是否在一定的前提条件下赋予其法律主体地位，并不存在必然的冲突，我们完全可以在协商制定人工智能体伦理规则与保障规则的基础上，赋予其一定的法律主体地位，从历史发展与人类整体角度看，这并不违背人类的利益。

刑法应赋予人工智能体以独立的被害者地位

人工智能体承载着人类的基本道德情感

从刑事立法角度看，刑法是由统治阶级将那些严重悖离社会道德、违反社会秩序的行为规定为犯罪的法律。Joel Feinberg 将损害原则和冒犯原则作为刑法犯罪化的完整道德基础，损害原则被认为是对人类福利性利益的破坏或妨碍，

而人类在各个阶段的生理健康、精神状态上的善好利益，以及人生的远大抱负，都是人类应该享有的福利性利益；而冒犯行为是引起他人不快的精神状态的行为，其具有导致损害结果的可能性，刑法只能对极其严重，且受众难以避免的冒犯行为进行规制（郑玉双，2016:185–86）。

冒犯行为因为直接与滋扰行为引起的人们的精神不快结合起来而被认为与情感有关，但本文认为，其实损害原则所依据的福利性利益，归根结柢也是情感的宣泄与表达。因为肉体的剧烈疼痛会伴随精神的伤害，他人生命被非法剥夺，会给其家人朋友带来痛苦与折磨，财产受侵夺、欺骗、被他人非法占有，也会让所有者因为失去物质保障而无助失落。某种意义上可以说，享有福利性利益的最终目的，是为了提升个体的精神状态，使情感有所依托并得到释放。其实，原始的人类社会规则就是建立在集体道德观上的，行为一旦触碰道德底线，就会遭受惩罚。正如英国社会人类学家 Alfred Radcliffe-Brown 所言，原始社会中，「一个社会中公认的不法行为……，其核心就是群体对因为内部成员侵犯了公认的群体道德观念而导致的社会动荡状态的反应。在这个反应中包含了集体道德愤怒的情感，从而起到使社会恢复安宁的作用。它的最终目的就是保持社区成员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 也认为，犯罪不是对权利的侵害，而是对基本道德情感的侵害（加罗法洛，1996:44）。随着理性的成长，人类社会才逐渐将惩罚的依据由集体道德外化为貌似更具客观性的利益，但从根本而言并没有完全摆脱集体道德观念的影子。可以说，不法行为引起利益的客观状态面受损，只是犯罪化正当性的表面依据，基于利益归根结柢是「人」的利益之理解，本文认为，犯罪化的实质正当性，仍在于抗制和打击严重破坏社会成员情感的

行为。对此，梁根林（2005:41）曾言，道德的基础，是将社会中的人假设为普遍善良的个体，并且以此作为社会治理的重要标准，以期社会中的普通人都能成为无我、忘我的天使，而刑法存在的基础，是将社会中的个体假设为性恶的个体，利用刑法的治理，就是利用刑法压制人性之恶。

未来，智能机器人不但将以深入社会的交往方式与人类进行情感交流，它们自身也完全可能产生与高级生命物一样的感觉和情绪。Alan Turing 的图灵测试（Turing Test）就强调了智能机器人除了要具备人类感知能力之外，也要具备能够与人类进行情感互动的能力，「在重视人工智能完成任务和功能强化的同时更要建立和满足人的情感和心理需求，这才是人工智能的最终定义。」（张爱萍，2016）

现实生活中也需要愈来愈多的能够与人类进行情感交流的智能机器人，比如伴侣机器人，它需要能够通过对人类面部表情、语言表达、肢体动作等外在表现进行情感计算和分析，从而读懂人类情感并表达自己的想法，满足我们对社交的内在渴望。为了实现人类的陪护需要，未来智能机器人也会相应地产生疼痛、难受、恐惧或快乐等现有的生命体所具有的感知力，以减少与人类的沟通障碍。当智能机器人也具有可以表达自己情绪的能力时，就有必要探讨是否应该将人类对智能机器人的某些行为纳入道德和法律的范畴。Ray Kurzweil 在《人工智能的未来：揭示人类思维的奥秘》（*How to Create a Mind: The Secret of Human Thought Revealed*）中说过：「当机器说出它们的感受和感知经验，而我们相信它们所说的是真的时，它们就真正成了有意识的人。」而大多数道德和法律制度也是建立在保护意识体的生存，和防止意识体受到不必要的伤害的基础上的（高奇琦，2018:29）。

与之相应，人类也会将重要的情感寄托在某些智能机器人身上。如日本的陪护机器人「帕罗」（PARO），可陪伴

老年人唱歌、跳舞、游戏等，颇受欢迎。为了获得更逼真的体验效果，愈来愈多的性爱机器人无论从外形上还是从内在情感互动上都逐渐趋近于人类，尤其是矽胶打造的皮肤和面庞，能够增加人类的亲切感，这种与人类贴身接触的机会，难免让人类对其产生依恋、愉悦的情感。而如果这些基本的情感遭到破坏，就容易对这些破坏行为产生出违背道德观念的罪恶感。2014 年，一位人机交互（human–computer interaction）专家做了一个实验，她让人类主动伤害形象真实可爱的机器人，然后记录人类的身体反应。实验后，大部分人类都表示这一行为让他们感觉到深深的不安，道德意识较强的实验者甚至产生了对于自我的较强的抵触意识（搜狐，2016）。

从现实角度看，人类对智能机器的情感最初表现为依赖感，即将个人决策建立在智能机器的分析信息之上。比如当前我们依赖智能导航系统提供的驾车路线行驶，医疗诊断依据大数据提供的概率进行判断，出行的意愿会受到智能算法得出的天气预报影响等。等到人工智能发展出情感互动等类人功能，人类原先基于机器人的决策功能产生的依赖感，会逐渐转变为更深层次的依恋感，并进而延伸出诸如怜悯、同情等情感。[5] 事实上，怜悯或同情是人类产生的一种由己及他的良善情感，这种情感并不仅限于对人类或动物，只要是一种与人存在情感互动之物处于被侵害状态，都有可能触发

5. 打个也许不恰当的比喻，这就像幼童对于父母感受的变化那样，一开始只是出于父母满足了自己基本的生活需求而产生依赖感，但是父母不断给予幼童情感的呵护，幼童就会逐渐将依赖感转变为依恋感，即使幼童长大不再依赖父母，也会基于依恋感与父母保持长久而亲密的关系，从而有了基本的家庭伦理道德。

人的怜悯、同情之心。智能机器人技术的发展，将使机器人从外形、思维方式和行为方式上都比动物更像人类，人类会更容易把智能机器人当作人来看（Scheutz, 2012:207），所以当智能机器人受到伤害时，人们会更容易产生怜悯和同情，从而让人对这些伤害行为发出道德情感上的谴责。例如，美国军方曾经让智能机器人踩踏地雷进行拆除的测试，但最终上校下命令终止了该测试，原因是每当该智能机器人踩到一个地雷，它就失去一条腿，并借助剩下的腿继续进行，「上校无法忍受看到被烧伤、伤痕累累和残废的机器用最后一条腿拖拽行动的痛苦，他控诉这项测试是不人道的。」（瑞恩•卡洛、迈克尔•弗鲁姆金、伊恩•克尔，2018:220）

本文认为，这种移情正是将侵犯人工智能体的行为予以犯罪化的伦理根基。基于刑法的道德性所应该与能够发挥作用的范围并不排斥非自然生命物，因此，如果刑法关注的是犯罪行为本身及其对他人的影响的话，那么，将对人类的犯罪行为施加到智能机器人身上，同样也应该被认为是犯罪行为。从刑法上重视这种道德情感的原因，还在于刑法的人道并非「对人」之道，而是「为人」之道，残酷本身除了会使民众对刑罚变得「麻木不仁」，使刑法的预防目的落空之外，还会松绑人的道德约束，从而增加犯罪行为发生的概率。有论者指出，如果法律没有对针对智能机器人的伤害虐待等行为进行保护，便极有可能为公众带来消极情绪，从而纵容人类的残暴行为，并最终使这些行为转化为针对他人的违法犯罪行为（刘宪权，2018:199）。总之，如果我们能够接受甚至力促将智能机器人作为刑法上的犯罪人主体，就没有理由否认其亦能够成为刑法上的被害主体。

智能机器人已经嵌入了人类社会秩序之中

刑法作为国家的统治工具，具有维护社会秩序稳定的功

能，在社会结构的调整过程中，这种功能被内化，进而成为犯罪化的正当性依据之一。事实上，自 1997 年中国的新刑法颁布以来，立法者基于社会治理的需要，从实现社会安全、秩序稳定和经济发展的立场出发，增加了很多反道德性并不明显的行政犯（时延安，2018）。在风险刑法和预防刑法理论盛行的当下，法益的提前保护为秩序维护这一刑法的正当性根据提供了更多的施展空间。应该说，刑法维护社会秩序的目的转向，使得对智能机器人的刑法保护更具可能性。原因在于智能机器人从某种程度上说，已经深深地嵌入了当下的社会秩序之中，承担了各种重要的社会职能，一旦其受损，遭到侵害的实则是其背后的社会秩序与社会机能。

具言之，当前智能机器人通过充当社会各个领域的工作者和活动的参与者，已经与社会秩序产生实质性的紧密联系。智能机器人加入的社会领域已经愈来愈广泛，高度智能化带来的优势使其不再局限于完成简单低级的工作，而是正在成为某些领域的精英。比如，将杀伤性军用机器人投入战争；自动驾驶机器人投入交通运营；智能写作机器人开始投入新闻、股评、诗歌乃至小说的创作；医疗机器人投入医用手术；家庭服务型机器人投入家政服务市场；在金融领域，智能机器人在速度和数据整合准确度上，已经逐渐超过金融分析师，现在纽约和伦敦证券交易所的交易大厅几乎形同虚设，真正的交易过程已经全面实现了「机器自动化」（中国日报网，2016）。未来，智能机器人甚至可能走进政治领域，成为办事高效的公务协助员，从事大量行政事务性工作，尤其是成为交通领域的机器人警察，运用电子眼提取并分析交通事故数据，极速开出罚单，提高交通事故处理效率，也能够以智能定位系统追踪可疑车辆，或者指示车辆行进畅通路段，大街小巷都可以看到能够经受日晒雨淋的电子交通巡警机器人。

将来智能机器人大范围参加社会各个领域的实践，导致它们成为维系社会秩序稳定和创造新秩序的重要一员。所以，未来对智能机器人的损毁、虐待、剽窃或窃取数据等行为，完全可能成为破坏社会秩序的严重行为。例如随意损毁或殴打在行政机关从事公务的智能机器人，已经危及了政府的管理秩序；为了毁灭证据而消除记录了刑事案件关键信息的智能机器人的数据，危及了司法秩序；剽窃智能机器人生成的具有独创性特徵的智力成果，破坏了著作权秩序；窃取并控制智能机器人金融分析的重要数据，威胁了市场金融秩序等。

之所以要重视这些被智能机器人参与的社会秩序，原因在于这些领域承载了人们的重要利益，在数据共享时代，人类的重要利益一般都属集体利益，而某一领域的秩序混乱，最终会使集体利益受到严重损害。此外，一个侵害智能机器人的不法行为除了导致社会秩序混乱，还会进一步引发其他犯罪现象。著名的「破窗理论」（broken windows theory）可以很好地揭示无序的环境对犯罪的影响。[6] 这里的无序不单是指物理环境的脏乱差，还包括恶劣的人际关系及越轨行为。这表明无序对人的越轨行为或者违法犯罪产生了强烈的暗示性或者诱导性，因为无序体现了某种程度上犯罪控制力

6. 破窗理论认为，无序「将使一个社区以螺旋形的方式慢慢失去控制，其中的居民也会渐渐躲避、退出或者逃跑；这种结果又反过来进一步加剧了社区中非正式控制机制的消失，并导致更为严重的犯罪，进而导致恐惧的增加，等等。随着社区的衰败，无序、恐惧以及犯罪螺旋式上升。」参见麦克•马圭尔等（2012:683）。

的薄弱（参见李伟，2014:137）。智能机器人全面嵌入社会领域后，任何一个破坏智能机器人正常活动的行为都有可能制造出社会的无序性，从而鼓励其他违法犯罪现象的发生。比如，剽窃智能写作机器人的智力成果的行为如果得不到法律的及时制止，就会让潜在的犯罪人看到获益的契机而肆无忌惮地剽窃智能机器人的智力成果，甚至会蛊惑更多守法公民剽窃他人的智力成果，最终扰乱著作权领域的秩序，削弱人们的创作热情，还变相鼓励了不劳而获的非诚信行为，导致减损人的整体道德感，从而又引发其他的犯罪行为。

由于未来智能机器人承载着人类的基本道德情感，与社会秩序产生愈来愈密切的联系，所以刑法有必要进一步考虑对智能机器人的保护，使其免受不法侵害，这也是发展人工智能道德的必要内容。最新的欧盟委员会「可信赖的人工智能道德准则草案」（Draft Ethics Guidelines for Trustworthy AI）指出，「我们必须确保最大化 AI 的优势，同时降低风险，因此需要以人为本的人工智能方法，AI 的开发和使用不应被视为一种手段，应视为增加人类福祉的目标。」（搜狐，2018）其将确保人工智能的「道德目的」作为可信赖（trustworthy）人工智能的组成要素之首，使人工智能为个人和社会的福祉而发展。但这种「道德目的」不应该仅表现为单向的人工智能对人类权利的尊重与规范的遵守，还应同时考虑人类对智能机器人相应的保护，将智能机器人的主体性考虑在内，要求自然人同样尊重智能机器人的某些重要利益或重要活动，减少人类对智能机器人的非人道的、无序的行为。从刑法的角度来看，未来刑法不仅会规制智能机器人的犯罪行为，也会打击针对智能机器人的犯罪行为。

受侵害人工智能体的刑法保护与救济路径

现阶段：借用行政犯的立法模式来保护人工智能体

当前，法定犯（statutory offence）时代的到来可谓是一种社会发展的必然趋势。应该说，法定犯的规定，并非像自然犯一样具有高度的民众认同度，大多数情况下，其是基于一定时期内维护社会秩序的需要而被立法者制定的。换言之，法定犯不是为了直接地保护某个具体受害人的个别利益免受损害，因为法定犯所规制的行为本身并不具有强烈的道德可谴责性，典型的法定犯立法模式往往规定了「违反……某行政法规的规定」之类的空白罪状，从而与行政法规产生了内在联系，应该说，不论是空白罪状中的行政法规还是法定犯本身，最终都是为了国家的管理制度能够有序运行而存在，其背后所维护的是超个体的集体法益。换言之，法定犯所保护的法益并非具象的人或物的状态，而是该状态背后的法秩序。

就智能机器人受侵害的问题而言，短时间内要让智能机器人获得刑法上的主体资格并不现实，但是，这并不影响刑法规制对智能机器人的侵害行为，原因在于智能机器人所嵌入的社会秩序与价值体系本身已经达到了刑法介入的程度，事实上也早已处于刑法所保护的范围。具言之，如前所述，智能机器人将承载人类愈来愈多的重要道德情感和重大利益，对智能机器人的某些破坏或干扰行为，将可能直接导致人类的这些重要利益受到损失，即使还没有获得法律主体资格的承认，刑法也可以将这部分侵害行为，以典型法定犯的形式专门规定下来，以避免对人类利益保护的不周延。

或许有人认为当前的罪名可以涵盖一部分通过侵害智能

机器人进而危害人类的行为，如故意破坏自动驾驶汽车的控制系统，使其在行驶过程中足以发生倾覆、毁坏的危险，可以认定行为人触犯了破坏交通工具罪，但是如前所述，随着自动驾驶技术智能化和自动化的提升，未来自动驾驶汽车不会被简单定义为交通工具，而是通过获得极大自主性而具备主体资格的自动驾驶机器人，因为其在极大程度上脱离程序控制独立驾驶发生交通事故后，完全可能由其独自承担法律责任，这将与传统意义上的以客体物为存在形式的「交通工具」产生本质性差异，即只有那些非智能或弱智能的火车、汽车、电车、船只和航空器，才会被解释为破坏交通工具罪中的「交通工具」。

此外，诸如当前已经出现的利用性爱机器人开设妓院的行为是否合法？若不合法，是否应当受到刑法追究？是否可以适用已有的组织卖淫罪、聚众淫乱罪等罪名对此行为进行规制？此等问题均存在较大争议。这说明，智能机器人的出现对传统刑法的理论与实务提出了较大挑战，刑法当前的罪名无法始终或准确涵盖针对智能机器人的侵害行为，忽视智能机器人技术的进步而坚守旧的规则，难以适应社会的新发展和新要求。

如果说，刑法以主动的姿态介入规制严重侵害智能机器人的行为是基于一种人本主义的功利性目的的话，则应首选法定犯的规定模式，因为一方面法定犯不过分关注其所直接保护的具象的人或物的状态的本然性质，从而有利于回避「智能机器人是否应具有主体资格」的争议；另一方面，就立法技巧而言，其规制模式具有便利性和灵活性，因而使得此类犯罪的规定更具有包容性，从而令刑法能够始终保持与科技发展的及时双向互动。最后，从反面看，如果法定犯的立法模式成立的话，势必意味着存在行政法的前置性保护，

这将使智能机器人及其所承载的重要人类利益受到双重保障。

综上所述，本文认为，可以先制定一部类似于《智能机器人管理与保护条例》的行政法规，以明确规定享有法律保护的智能机器人的定义，如「本条例所称『机器人』是指以服务社会为目的，能够独立行为，具备良好的类人性的感知和控制系统，可以通过大数据处理进行深度学习和分析的类人型智能机器人。」并规定不得出现针对智能机器人的各项不良行为，例如「不得随意肢解、损毁智能机器人；不得非法干扰智能机器人正常工作；不得利用智能机器人做违背公序良俗的事情……」等，并以罚款、拘留等形式确保该条例得到贯彻落实。进而《中华人民共和国刑法》可以在总则第五章「其他规定」中，增加刑法中智能机器人的定义，具体可以参照《智能机器人管理与保护条例》的行政法规中对智能机器人的定义内容；在分则中设立专章，规定有关智能机器人受侵害的犯罪，可以在具体条文中，均以「违反保护智能机器人的规定」为前提条件，辅之以行为要件和情节要件，作为该条文的要件内容，例如，「违反保护智能机器人的规定，公然与智能机器人进行性交，情节严重的，处……；情节特别严重的，处……。」当然，就这些新罪名的法定刑配置而言，究竟有没有必要参考针对人类类似侵害行为的刑罚类型和幅度，还是无需参考，而径直设置这一类行为的刑罚域，还有待进一步商榷。

未来展望：直面人工智能体自身法益的保护模式

如果说上述现实保护模式是以保护人工智能体所嵌入的社会秩序为基底的间接模式的话，那么，在未来，智能机器人的主体地位获得法律的认同之后，刑法可以抛开前置性的

行政规范，径直对人工智能体展开保护。具言之，本文认为，智能机器人理应获得法律上的主体身分认同，其自身的利益诉求，也会随着法律主体资格的确认而被刑法所重视，成为独立于人类的利益。姑且将智能机器人的那部分自身利益，即那些类似于人的生存权一样，能够维持智能机器人持续地存在、运作（以后或许称为工作）的利益，包括前文论述的不受无端破坏、肢解等，有独立作出判断、选择和创作的自由，以及收集、保有重要数据等利益，称为智能机器人的核心利益。本文认为，智能机器人保护的非人本主义利益观，要求我们将智能机器人的核心利益纳入道德和价值范畴，而不再将破坏智能机器人核心利益的行为视为破坏人类利益之附属的行为，即应该对这类危害性的行为予以独立的价值评价，使其成为刑法上可以脱离于人类利益而评价的对象。

Raffaele Garofalo 将典型的犯罪（自然犯）认定为是对人类怜悯情操和正直情操的违反（参见加罗法洛，1996:44），是一种显而易见的罪恶，如果将智能机器人的重要利益评价为独立于人类的重要利益，那显然可以将损害这部分利益的行为视为一种显而易见的罪恶。也即智能机器人获得法律主体地位的承认后，刑法的人本主义利益观才会彻底改变，智能机器人才会摆脱人类的附属物性质，真正以犯罪者和犯罪受害者的角色登上刑法舞台。此时，保护智能机器人切身利益的驱动力，从原来的以人类利益为中心的功利性需求，转变为发自内在的道德请求，那些挣脱道德约束而严重侵害智能机器人切身利益的行为，因为彻底违背主流道德观而受到刑法的谴责，此时，如肢解智能机器人并使之完全丧失行为能力的行为，并非因为使人类失去了重要的劳动工具而被规定为类似财产性质的犯罪行为，而是出于该行为与残忍地杀害人类的行为无异，是具有人身损害性的犯罪行为。

对此，刑法应重新调整侵害智能机器人的犯罪规

定——在前述法定犯模式的基础上，将侵害获得主体资格的智能机器人之核心利益的行为独立出来，并在《中华人民共和国刑法》总则第五章关于智能机器人的定义后面，增加其获得法律主体资格的标准，即在民政部门或科技部门获得类公民主体身分而有备案登记的类人型智能机器人。被单独分离出来的这部分犯罪行为根据法益类型，可以在《中华人民共和国刑法》分则有关智能机器人犯罪专章中，具体分为破坏智能机器人完整性的犯罪、干扰智能机器人行动自由的犯罪，和严重阻挠智能机器人收集、保留重要数据，修改、窃取其重要数据犯罪等三种类罪名。同时，对于那些没有直接危害到智能机器人核心利益的其他侵害行为，如剽窃创作型机器人的音乐作品，并以此获取数额较大的收益，由于直接威胁或损害到的是著作权秩序下的人类的利益，刑法仍然应该以法定犯的形式予以规定。

当然，随着智能机器人技术的发展，智能机器人的核心利益类型还会不断扩张，比如未来在智能机器人与人类能够友好共处的前提下，如果最终允许人类与智能机器人登记「结婚」的话，那么，也可能会产生智能机器人的婚姻自由；如果智能机器人有相互交往的需求的话，那么，势必会产生通信自由；如果允许智能机器人参与人类社会的管理或智能机器人群体内部的管理的话，那么，很可能会延伸出智能机器人的政治选举自由、信仰自由等。此时，对智能机器人刑法保护的正当性依据，可能更多地是取决于对行为本身的客观评判，而非立法者基于人本主义的考量（白建军，2018）。

本文认为，这时刑法势必也会相应地逐渐增加新的罪名规定，保护智能机器人的罪名类型，将会出现藉由保护社会秩序来保护智能机器人的间接保护模式，向直接保护模式拓

展的趋势。此外，侵害智能机器人核心利益的犯罪行为，还因为智能机器人嵌入社会秩序和价值体系后，与人类的利益相互牵连，因而表现出复杂的社会危害性，刑法对于这部分犯罪行为的打击态势也会更为严厉。

对智能机器人遭遇刑事被害的补偿与救济

法谚云：无救济，则无权利。对未来获得主体地位智能机器人的利益保护，应该比照对人类权利的保护，除了应该尽可能将智能机器人的利益纳入刑法和其他法律范畴予以事前保护外，也应该考虑对那些受犯罪侵害的智能机器人的利益进行事后救济。Christopher Stone 提出，某一主体能否拥有法律权利应满足以下条件：第一，该主体应其要求可以提起法律诉讼；第二，法院在决定授予法律救济时必须考虑到损害；第三，法律救济必须满足它的利益需求（McNally and Inayatullah, 1988:126）。

在恢复性司法（restorative justice）理念下，现有的刑事犯罪所要考虑的救济问题，着重于对被害人的赔偿问题，并认为刑事损害赔偿不仅能使被害人的利益获得实质性的保护，对预防犯罪也有一定作用。[7] 刑事赔偿包括物质赔偿和非物质赔偿，当前关于被害人赔偿的主体包括犯罪人赔偿和国家赔偿。在中国，针对前者，主要通过被害人或其法定代理人、近亲属提起附带民事诉讼的方式，要求犯罪人赔偿，

---

7. 美国全国少年司法研究中心（National Center for Juvenile Justice）曾在犹他州调查了6,336件官方统计的少年假释案件。结果发现，赔偿的使用与一些少年犯罪人中累犯的行为的显著减少有正面联系（U.S. Department of Justice, 1992:4; 李伟，2014:208）。

而智能机器人脱离于自然人属性，加上侵害行为往往可能严重损坏智能机器人的智能感知系统，所以其独立提出附带民事诉讼的能力不足，对此，检方或与智能机器人关系密切的其他公民，可以根据智能机器人的修复情况，代替其决定是否提出附带民事诉讼请求，具体的赔偿金额，可以要求犯罪人返还对智能机器人剥夺的经济利益，或者参考修复被损智能机器人的费用，以及日后加大保养的费用，由智能机器人的所有者或管理者代为保管使用；亦或要求犯罪人在限期内重新修复智能机器人、给智能机器人赔礼道歉等。此外，国家也应该就某些不当的国家行为，对智能机器人造成的重大利益损害进行赔偿，尤其是涉及到对智能机器人本身与财产性的利益损害时，应当参照自然人赔偿标准，对智能机器人进行弥补。

余论与展望

本文所讨论的主旨问题，是刑法究竟应该如何面对和保护人工智能体，本文的基本观点是人工智能体最终应该具有独立的被害人地位。但其实真正让人工智能体能够获得法律主体性承认的，一方面是人工智能体能否在未来社会中发挥愈来愈重要的正面积极作用；另一方面，更深层次的是人类能否突破人类中心主义思想，对「自我」与「他者」之间的关系进行更为深刻的反思。

应该说，现有的刑法理论是在人本主义层面上建立起来的。人类经过漫长的探索，才摆脱神灵的控制，进而认识到人自身的意义与价值，人是自由的，这意味着人是自己的主宰，社会定纷止争的权力应该是保护人，并由人来制定与执

行的，而刑法是执行人类意志最强有力的规范武器，所以刑法一开始就是由人所制定，并为人类服务的法律。人工智能体作为人类的生成物，是与人类不同的「他者」，是现有刑法制定主体与调整对象之外的存在物，站在人类中心主义立场看，除非人工智能体被视为与人类财产一样的附属物，否则其几乎无法与刑法有任何实质的交集。然而事实并非如此，人类中心主义思想已经在日益严重的环境污染和生态破坏的现实情形下暴露出了自身的弊端与局限性，当下人工智能体的社会属性早已超越了其自然属性，从关系本体论出发，人类的整体幸福感未必不可以建立在与自身以外之存在物的广泛交互性上，这必然促使人类更加注重「他者」对于社会价值与功能的发挥。刑法也应该维护人工智能体这种「他者」的社会价值，如同环境对人类的价值需要，由刑法对那些破坏生态平衡的严重危害行为施以惩罚来保护一样。

当前刑法面临抉择之处就在于能否以及如何突破人类中心主义的立场，给人工智能体这样的非人类存在物之「他者」一个庇护之所。或许古典功利主义理论（utilitarianism）能提供一种现实的解释途径。因为功利主义理论仍然以人类中心主义为基础的，只要能够「实现最大多数人的最大幸福」，就是值得肯定的，以此为目标建立的制度就是合适的制度，保护人工智能体可以被解释为是为了维护人类的最大利益，可以提升人类的整体福利，所以刑法保护人工智能体是合适的。这样刑法就可以在保有原来立场的同时，也能给予人工智能体一定程度的制度保护。但没有突破人类中心主义立场，就意味着对人工智能体的保护受限于人类的需求与利益，只要不涉及人类的利益，人工智能体就不在刑法的保护范围内，这样与其说是保护人工智能体，不如说是在保护人类自身，进而「自我」与「他者」之间的矛盾仍然没有得到调解。

本文最后展望，未来的刑法会突破人本主义的制约，将人工智能体这样的「他者」当作人类自身一样的存在物去对待，这既是人类自我观念转变的结果，也是人类文明进一步开化的结果。届时刑法将既保护人类自身，又平等地保护诸如环境、动物和人工智能体一样的非人类存在物。

参考书目

A. R. 拉德克利夫 • 布朗。2014。《原始社会结构与功能》，丁国勇译。南昌：江西教育出版社。

中国日报网。2016。〈全球金融业进入『机器人时代』：你的血汗钱是否安全？〉，3 月 21 日。取自：http://m.cankaoxiaoxi.com/finance/20160321/1105779.shtml。

王肃之。2018。〈人工智能犯罪的理论与立法问题初探〉，《大连理工大学学报（社会科学版）》，第 39 卷，第 4 期，页 53–63。

王耀彬。2019。〈类人型人工智能实体的刑事责任主体资格审视〉，《西安交通大学学报（社会科学版）》，第 39 卷，第 1 期，页 138–44。

加罗法洛。1996。《犯罪学》，耿伟、王新译。北京：中国大百科全书出版社。

司晓、曹建峰。2017。〈论人工智能的民事责任：以自动驾驶汽车和智能机器人为切入点〉，《法律科学（西北政法大学学报）》，第 5 期，页 166–73。

白建军。2018。〈法定犯正当性研究：从自然犯与法定犯比较的角度展开〉，《政治与法律》，第 6 期，页 2–12。

皮勇。2018。〈人工智能刑事法治的基本问题〉，《比较法研究》，第 5 期，页 149–66。

刘宪权编。2018。《人工智能：刑法的时代挑战》。上海：上海人民出版社。

托比•沃尔什。2018。《人工智能会取代人类吗？》，闾佳译。北京：北京联合出版公司。

江山。2000。〈法律革命：从传统到超现代——兼谈环境资源法的法理问题〉，《比较法研究》，第 1 期，页 1–37。

张玉洁。2017。〈论人工智能时代的机器人权利及其风险规制〉，《东方法学》，第 6 期，页 56–66。

张爱萍。2016。〈与机器人谈『感情』，人类是否『很受伤』？——也谈人工智能与人类情感融合的前景〉，光明网，4 月 5 日。取自：http://m.cankaoxiaoxi.com/science/20160405/1118962.shtml。

时延安。2018。〈犯罪化与惩罚体系的完善〉，《中国社会科学》，第 10 期，页 102–25，206–07。

李伟编。2014。《犯罪被害人学教程》。北京：北京大学出版社。

杜严勇。2014。〈现代军用机器人的伦理困境〉，《伦理学研究》，第 5 期，页 98–102。

麦克•马圭尔、罗德•摩根、罗伯特•赖纳等。2012。《牛津犯罪学指南》，刘仁文等译。北京：中国人民公安大学出版社。

阿西莫夫。2005。《机器人短篇全集》，汉声杂志译。北京：天地出版社。

郑玉双。2016。〈为犯罪化寻找道德根基：评范伯格的《刑法的道德界限》〉，《政法论坛》，第 34 卷，第 2 期，页 183–91。

高奇琦。2018。《人工智能：驯服赛维坦》。上海：上海交通大学出版社。

曹明德。2002。〈法律生态化趋势初探〉，《现代法学》，第 24 卷，第 2 期，页 114–23。

曹明德。2007。《生态法新探》，第 2 版。北京：人民出版社。

梁根林。2005。《刑事法网：扩张与限缩》。北京：法律出版社。

焦艳鹏。2012。《刑法生态法益论》。北京：中国政法大学出版社。

搜狐。2016。〈《西部世界》这部 9.2 分的『小黄片』凭什么封神〉，11 月 3 日。取自：https://www.sohu.com/a/118039969_520625。

搜狐。2018。〈刚刚，欧盟 AI 道德准则草案出炉！可信赖的 AI 才能成为人类的北极星〉，12 月 19 日。取自：https://www.sohu.com/a/282938991_354973。

瑞恩 • 卡洛、迈克尔 • 弗鲁姆金、伊恩 • 克尔编。2018。《人工智能与法律的对话》，陈吉栋、董惠敏、杭颖颖译。上海：上海人民出版社。

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.

被害者视角

刑法何以保护人工智能体？

摘要

人工智能体在广泛地参与人类社会生活过程中，其独特的利益诉求、与生俱来的利他性，以及未来独立承担法律责任的可能性，使得其理应获得类似于人类主体的法律主体地位。人工智能体将承载愈来愈多的人类道德情感，并与社会秩序产生日益紧密的联系，这些特徵强化了其与刑法的内在联系。刑法应在主体性视角下保护人工智能体自身及其承载的重要利益，应将对人工智能体的某些非道德或无序行为纳入犯罪范畴。当前可先以典型法定犯的模式规定此类犯罪，待未来人工智能体的法律主体地位得到普遍承认时，再逐渐向直接保护的模式转型，并应在恢复性司法理念支配下，对受侵害的人工智能体予以必要救济和补偿。

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