Open Journal of Legal Science 法学, 2023, 11(1), 7-12
Published Online January 2023 in Hans. https://www.hanspub.org/journal/ojls
https://doi.org/10.12677/ojls.2023.111002

Hans汉斯

论人工智能刑事主体地位之否定

薛荷蓉，苏贤桂

新疆大学，新疆 乌鲁木齐

收稿日期：2022年11月4日；录用日期：2022年11月15日；发布日期：2023年1月4日

摘　要

对于是否应当赋予人工智能刑事责任主体地位在学者间展开了论战。本文从赋予人工智能刑事责任主体地位的前提、资格、目的与风险四个方面进行审视，认为是否能够真正实现强人工智能尚且无法知晓、其意志不具有社会规范属性、对其实施刑罚无法达到刑罚目的以及存在颠覆人类主体地位、不受约束等风险，其刑事主体地位不适格。涉及人工智能的犯罪行为所主要发生的财务管理、自动驾驶及医疗领域的刑事责任的承担需视情况而定。

关键词

人工智能，刑事责任主体，刑罚

On the Negation of the Criminal Subject Status of Artificial Intelligence

Herong Xue, Xiangui Su

Xinjiang University, Urumqi Xinjiang

Received: Nov. 4th, 2022; accepted: Nov. 15th, 2022; published: Jan. 4th, 2023

Abstract

There is a debate among scholars on whether artificial intelligence should be given the status of subject of criminal responsibility. This paper examines the premise, qualification, purpose and risk of giving artificial intelligence the status of subject of criminal responsibility, and believes that whether it can truly realize strong artificial intelligence is still unknown, its will does not have the attribute of social norms, its punishment cannot achieve the purpose of punishment, and there are risks of subverting the human subject status and not being constrained, and its criminal subject status is not eligible. The criminal responsibility in the fields of financial management, automatic driving and medical care, which mainly occur in criminal acts involving artificial intelligence, shall depend on the situation.

文章引用: 薛荷蓉, 苏贤桂. 论人工智能刑事主体地位之否定[J]. 法学, 2023, 11(1): 7-12.
DOI: 10.12677/ojls.2023.111002

Keywords

Artificial Intelligence, Subject of Criminal Responsibility, Penalty

1. 问题缘起

科学技术在人类历史的演变和发展过程中有着摧枯拉朽的力量和优势。随着人工智能进入大众的视野并不断普及和升级，人工智能(AI)已涉及人们生活的方方面面。然而，人工智能给我们的生活方式带来颠覆性影响的同时，也对传统法律发起了更多挑战。本文就人工智能能否成为刑事责任主体问题进行讨论。

人工智能是指使机器像人一样去完成某项任务的软硬件技术。[1]以伊隆•马斯克提出的“AI 威胁论”为代表，人工智能威胁论越来越多的被提及。相反，以扎克伯格为代表的部分业界人士则认为人工智能威胁到人类生存与发展之根本的时代尚且遥远，并且该时代是否真正会到来也尚未可知。对人工智能不同层级的理解导致这两种观念的产生。伊隆•马斯克所言的人工智能是能够以人的自由意志进行思考并支配行动，具有处理诸多不同种类的任务和对于事先未曾预见的突发事件适应与应对能力的强人工智能。而扎克伯格针对的则是较为侠义的具备某一领域专业技能的有着明显的工具性的弱人工智能。我们目前所处的，并且未来很长一段时间都将继续处于弱人工智能时代。目前对于强人工智能只是一种期许，但是依据人工智能科学研究界理论来说，只要突破目前的技术障碍，强人工智能时代的到来指日可待。

正是基于这样的强人工智能的设想和理论研究，引发了刑法学者们对于强人工智能超越人类智慧、挣脱数据、代码、程序限制“独立行走”、以自主意识支配独立实施行为的忧虑与刑法规制探索与思考。因此，能够超越人类预先设定的程序范围，做出自我决策的强人工智能才是刑法所研究的对象，也是本文所讨论的对象。

2. 涉人工智能刑事责任主体地位的问题论战

对于强人工智能能否成为刑事责任的主体，以刘宪权教授为代表的肯定论者与刘艳红教授为代表的否定论者展开了理论博弈。

2.1. 否定论之“反智化”批判

否定论，即否定强人工智能的刑事责任主体地位。针对得益于国家政策信号而热衷甚至膨胀的人工智能刑事责任理论研究现象，刘艳红教授撰写《人工智能法学研究的反智化批判》一文不点名对该现象与相关学说理论进行批判。其主要观点为：其一，目前的人工智能刑事主体地位研究热潮，是基于一些学者混淆了机器机械化操作与刑事主体所要求的独立意志与行动，从而制造出的学术泡沫。其二，被作为典型论据的，机器人索菲亚获得公民身份不过是国家、企业与媒体行业的联袂炒作、营销手段。由此引发的人工智能法律人格问题实质是伪命题。其三，目前学界对人工智能的研究问题泛滥但体系不协调，其价值有待商榷[2]。加之，肯定论者所提出的刑罚措施，不能使没有共情力与同理心的人工智能感受到痛苦、畏惧等，因而也无法实现预防犯罪的目的，这样的刑罚毫无价值。

2.2. 肯定论之“伪批判”

刘宪权教授以《人工智能法学研究“伪批判”的回应》与刘艳红教授隔空对话。其一，他认为批判者忽略了人工智能所具备的在一定程度上大体大脑功能的特性，双方的讨论缺乏同一概念基础。其二，对方未区分人工智能不同发展阶段而以偏概全、移花接木将讨论对象偷换。在智能科技跨越式发展以及风险社会背景下，强人工智能并非学者们闭眼塞目的主观臆想。其研究具有现实意义，刑法本身也应当具备一定的前瞻性。其三，目前针对强人工智能主题所提出的处罚方式并非刑罚，也不排除强人工智能时代真正到来时重新制定与之相适应的刑罚制度[3]。

在批判与反批判论战中我们可以看出双方争议的焦点仍可归于是否应当赋予人工智能刑事主体地位。肯定论认为强人工智能是脱离了人类意志的体现或从属，拥有独立的意志，能够辨认和控制自己的行为并且能够独立的对行为做造成的不利后果承担责任。而否定论则与之相反。本文从前提、资格、目的、风险四方面对人工智能的刑事责任主体地位进行审视，认为否定论应是最优选择。

3. 人工智能刑事责任主体地位的审视

理论学说的观点终究是要服务于司法实践。人工智能能否被赋予刑事责任主体地位资格也应当充分考虑实际需要。强人工智能是否具有刑法意义上的对于自身实施的能够产生社会危害性的行为的认识和控制能力以及对其处以刑事处罚能否达到预期的处罚效果，实现刑法的目的是能否赋予其刑事主体地位的关键所在。再者，依据强人工智能的定义来看，它的存在本身就是对社会人类主体地位的极大威胁。人工智能能否依照刑事法律规范进行活动，这些规范对其是否具有约束力，是否有对人类法秩序造成严重破坏的高度风险也需进行探讨。

3.1. 前提审视

就客观事实来说，弱人工智能虽然对我们的生产生活方式产生了颠覆性的影响，其应用范围也不断扩大。但目前仍停留在深度学习的发展阶段，其发展速度与核心技术进展之间的差距不容小觑，难以实现质的飞跃。就拿人工智能领域最现实和前沿的，通过对于人的大脑结构以及大脑神经元的模拟以期达到实现强人工智能的“智能”研究来说，谷歌大力支持、耗时长久的一万六千多个处理器的研发，相较于人脑内动辄数十亿个神经元来说可谓沧海一粟。对于强人工智能所需的智能算法或是工程路线目前都是该领域的盲区，无从得知，其发展寒冬可能随时会到来。相当一部分专家对于所谓强人工智能的到来持怀疑态度。刘艳红教授所说的目前对于超强人工智能的研究完全是跨越时代的、诸多前提性条件尚未深入探讨的观点较为合理。肯定论者的观点前提始终还是一种假设，无实质可靠根据的理论如空中楼阁。

其次，从哲学的角度来看，所谓心智是理性和情感共同作用下的产物，在哲学文化发展和演变的数千年以来对于心智的定义都无法达成一个通识的概念，其复杂程度可见一斑。对于人类的心智的定义尚且如此，又如何以标准化、规范化的法律去界定人工智能是否具有自主意识、独立意志和智能程度呢？再者，我们所探讨的人工智能多是从一个理性人的角度出发而忽略了对情感的考虑和关注。但是人类所实施危害社会的犯罪行为是参杂了个人情感的结果。因此，缺乏情感但能够与人在理解、共情能力、智能方面高度契合的强人工智能体的出现甚至实施犯罪难以令人信服。

3.2. 资格审视

一个合格的刑事责任主体，首先要具备自由意志。其次，对于自己独立实施或者伙同其他主体实施的刑法所规制的危害行为必须具有辨认和控制能力。独立意志是辨认和控制能力的基础和前提，辨认和控制能力是独立意志的主要表现形式。就自由意志来说，肯定论观点所认为的，人工智能超越程序设定

所表现出的意志是其独立意志。然其并不具有刑法所要求的社会规范属性。人的独立意志不仅是人独立进行思考和决策，还包括对所学所感和所悟的反思。作为具有社会能动性的人对于自己的行为的善与恶、好与坏自有标尺。但人工智能所表现出来的意志需要我们站在第三人角度评价，不被社会承认与接纳，亦无法认定为刑法意义上的独立意志。

其次，就辨认和控制能力来说，强人工智能何以进行具有社会规范性的价值评价和判断呢？我们首先可以想到的是从人类设计、编程时向人工智能输入的伦理、道德、法律规范等可以作为其价值判断的来源，但如此一来则意味着其仍具有工具属性受控于人。倘若我们能以“道德嵌入”的方式在程序安排上使得人工智能具有价值判断的能力，那我们同样也能提前预测和评估危害行为发生的风险并将避免该风险的数字代码、工程路径等输入人工智能。那人工智能也不会产生违背道德或法律的意图，自然也就无需将其纳入刑事主体范畴。[4]或者，人工智能也可以凭借自身的强大的学习能力，通过不断地搜集学习、进化升级而具备人类的价值判断能力。这完全是脱离人类对其的设定而自主提升的结果。然而目前尚无理论依据证实其可行。就算其能够独立形成价值判断标准，该标准是否符合人类社会的规范也未可知。况且强人工智能完全可能通过自主学习和研发，形成一套符合人工智能体利益的价值判断标准。就像人类社会所形成的伦理、道德、法律等规范终究是为自身利益服务一样，届时刑法规范对其毫无约束意义。

3.3. 目的审视

对犯罪主体处以刑罚带来刑罚效果是其目的所在。肯定论学者为人工智能主体设置了修改程序、删除程序、永久销毁等刑罚方式，可能具有一定的合理性，但也伴随着致命的缺陷。首先，无论人工智能怎样发展，其本身是有别于自然人的无生命体，也不具有自然人的复杂情感，对其施加刑罚并不会让它在身体上或者精神上感受到痛苦和折磨，没有意识到这是对其行为的非难。因而也无法达到教育以预防犯罪的目的。其次，退一步说，即使未来的强人工智能能够因施加刑罚而感受到痛苦，这样的刑罚设置也难以达到刑罚的目的。因为拥有自由意志的人工智能不会受制于人。强人工智能本身具有较强的学习和提升能力，被删除、修改的程序完全可以通过自身加以修复，甚至如果删除、修改的次数较多，其还会对这样的刑罚产生“免疫”，自主摸索形成一套更具危害性的程序。即使删除、修改程序能使强人工智能感受到痛苦，那也是极为短暂、不值一提的，不会达到预防性目的。再者，即使是对它处以最为严格的刑罚—永久销毁，也难以达到像是对自然人处以死刑的效果。因为人工智能的本质是一套数学算法或说是工程路径，我们销毁的只能是人工智能体的容器、载体而不是其本身。它可以换一个载体继续存在，也即可以随时“复活”。因而永久销毁也变得毫无意义，只是徒增财产损失。肯定论者所提出的这些措施，从效用上来说最终可能沦为“机器维修手册”罢了。

3.4. 风险审视

如果强人工智能时代真的到来，世上就会广泛存在着能够和人类一样独立思考，有一套自己的行为标准和社会规范的，在各方面的能力，如记忆力、视力、听力、学习、思考能力以及运动能力都使人类望尘莫及的人工智能体存在。那将彻底颠覆人类的社会主体地位，我们可能成为被主宰的对象。把人工智能主体纳入刑事责任主体地位也是天方夜谭、痴人说梦，因为刑法对它们根本不会产生约束力。刑法的作用范围仅限于能够服从其规制的主体，如果强人工智能摆脱人类的控制束缚，能解决人类与强人工智能体之间的争端的最终只能是战争而并非法律。也就是说，真正的强人工智能时代的人工智能主体资格的学术探究终会沦为一纸空谈，因为其本身是个伪命题。再者，刘宪权教授的主张本身也存在着自相矛盾之处。如其主张强人工智能可靠自己的意志独立作为或不作为，可能会超越人类甚至奴役人类。又主张其法律地位的决定权仍在人类，应通过立法给其栓缰套绳。倘若强人工智能真实现了独立自主的行

动，人类制定的法律岂不是一纸空文？又何来对强人工智能的约束力呢？

综合以上四点的分析，本文认为人工智能不能也不应当成为刑事责任的主体。现代社会的分工逐渐明确、细化，在各方利益角逐中责任的承担都唯恐避之不及，因而也就会以各种形式来逃避或转嫁责任。从某种意义上来说，将人工智能归入刑事责任主体何尝不是一种转嫁责任的方式。人工智能既是人所创造出来的非生命体，创造者或使用人却不承担责任将会导致个人过于宽泛的自由，不利于约束个人行为，终究不利于社会进步。

4. 人工智能主体地位否定论下的实践展开

未来的强人工智能时代是否会到来我们尚不能准确预测。与人工智能是否为适格的刑事责任主体紧密相连的另一问题是涉及人工智能犯罪行为的刑事责任的分配，针对这个问题的分析更具有现实意义。为此，对于人工智能目前广泛应用的主要领域，即财产侵犯、汽车自动驾驶、医疗等人犯罪行为的定性与刑事责任分配是涉人工智能刑事责任研究主要应解决的问题。

4.1. 否定论下人工智能侵犯财产犯罪的定性

当下越来越多的产物管理工作在程序上都有人工智能的参与，如手机网上银行、支付宝、微信等第三方支付平台、淘宝等购物平台以及超市自动收银柜都是人工智能财务管理平台。而依然坚持机器不能成为被欺骗对象、须具有自然人的处分意思要素来严格区分诈骗罪与盗窃罪已显然不合时宜且流于形式。通过研究德日刑法我们会发现，在其法律中财产性利益未被归为盗窃罪的对象，因而当犯罪行为的作用对象为财产性利益时处分意识决定了其是否为犯罪行为。但是在我国承认财产性利益为盗窃罪对象的前提下，以这样的方式严格区分盗窃罪与诈骗罪的意义实际上有限。故而可无需固守处分意思的要件严格区分盗窃罪域诈骗罪，可将盗窃罪作为侵犯财产类犯罪行为兜底性罪名，在处分意思是否具备难以界定时认定为盗窃罪。[5]

4.2. 否定论下人自动驾驶汽车肇事罪责归属

根据汽车驾驶的自动化程度，我们可将其分为有条件驾驶、高度自动驾驶与完全自动驾驶三个等级[1]。有条件自动驾驶的汽车会在发生系统不能自动排出的障碍或做出选择时提示驾驶人，驾驶人此时需接管驾驶系统。其智能化程度较低，只要还是依靠驾驶人员的判断与操作。如发生事故应以交通肇事罪的标准来衡量。高度自动驾驶智能化程度较高，驾驶员的参与度低。应当依据驾驶人员有无主观罪过来判断罪责，无过失则判定为意外事件，但可基于设计或制造的缺陷追究设计人或制造人的责任。对于完全自动驾驶的汽车，其完全的智能化解放了驾驶员的手脚。该项技术的发展与运用利大于弊，顺应社会发展的需要，可以将其发生事故的风险看作是道路交通安全中“被允许的风险”。驾驶人或使用人因基于对该自动驾驶系统的信赖而产生的风险可以被接受。对于事故受害人的安抚，更多的应当是物质上的赔偿而非刑罚。因而可以通过汽车投保的或是导致产品缺陷的设计者、制造商方面取得赔付。

4.3. 否定论下医疗领域涉人工智能责任分配

目前，人工智能也被广泛的引入医疗领域，在疾病的诊断与治疗方面发挥着重要的作用。人工智能运用人类望尘莫及的数据存贮、运算、分析与识别能力促进了现代医疗事业的不断精进。但是，这些医疗人工智能不具有所谓的自主意识，其所作出的判断不过是基于预先输入的程序性代码或数据而做出的

[1]2018 年 2 月，上海市公安局等部门联合颁布的《上海市智能网联汽车道路测试管理办法(试行)》规定自动驾驶分三个等级：由系统负责全部驾驶工作，驾驶者需要在系统提示时接管驾驶，即“有条件自动驾驶”；由系统负责全部驾驶工作，即便系统出现提示，驾驶者也不必响应，系统能够自动响应并继续操控车辆安全行驶，即“高度自动驾驶”；由系统负责全部驾驶工作，车上可以没有乘客或者有乘客却不需要具备驾驶能力，因为不论在何种情境下都不需要驾驶者进行操作，即“完全自动驾驶”。

适当反应，其实质上是一种医疗器材。应当作为医疗领域内医疗机构或医务人员开展诊断、治疗活动的辅助性工具，不可本末倒置。若盲目依赖人工智能系统做出的判断而导致误诊造成医疗事故的，应当由医务人员承担刑事责任。而由于人工智能本身的设计或制造本身存在不被允许的缺陷和风险，则应当追究设计者或制造者的责任。

诚然，科研活动本身就是一种风险性活动，常伴随着财产损失风险甚至伤亡风险，人工智能本身的不确定性不要求设计、编程人员站在上帝视觉审视该技术，如果是现有科技无法发现也无法预测的危害，则刑法无需苛责相关人员，因为完全不具有期待可能性。[6]但若因设计、编程不周，或是进行技术测试时未按照行业内的规范标准进行完整测试亦或是其他过失造成严重事故，则应当承担刑事责任。

5. 结语

我们生产工具的目的在于工具能带我们走得更远。人工智能的到来开创了智能工具新时代。但由目前的弱人工智能过渡到预想中的强人工智能存在着难以跨越的技术鸿沟，即便能实现强人工智能，其自主意识与认知也无法具有刑法意义上的社会规范属性，以目前主流观点预设的刑罚措施对其进行处罚徒增财产损失而无法达到刑罚目的。再者，强人工智能意味着巨大风险的伴随，其一旦以自主意识独立行为，有颠覆人类主体地位之嫌，不能遵循人类社会的规范与准则。由此，人工智能并不适宜作为刑事责任主体之一。然而由人工智能引发的主要集中于财产、公共交通与医疗行业中的损害确需明确责任。就侵害财产来说，应当摒弃传统以认识、处分意识来严格区分盗窃罪与诈骗罪的标准，当处分意识难以界定时以盗窃罪作为适格罪名加以处分。在公共交通领域，应当以汽车的智能化程度、驾驶人员的参与度来划定罪责。对于参与度较高的有条件自动驾驶只需以常规的交通肇事罪标准进行判定；对于智能化程度高，驾驶人员参与度低的高度自动驾驶则以驾驶人员的主观罪过形式来判定；对于完全自动驾驶的汽车，则由于驾驶人员没有参与驾驶而无需对其追责。当然，设计者与制造者也需承担因设计、制造缺陷所导致的刑事责任。而在医疗领域中，人工智能只是辅助性的工具，医务人员应当作为承担医疗事故的责任主体。此外，科研的性质决定了每项技术的研发到逐渐成熟都要承担一定的风险，刑法对科研主体不能过于严苛。

参考文献

[1] Arruda, A. (2016) An Ethical Obligation to Use Artificial Intelligence? An Examination of the Use of Artificial Intelligence in Law and the Model Rules of Professional Responsibility. *American Journal of Trial Advocacy*, **40**, 443-458.

[2] 刘艳红. 人工智能法学研究的反智化批判[J]. 东方法学, 2019(5): 119-126.

[3] 刘宪权. 对人工智能法学研究“伪批判”的回应[J]. 法学, 2020(1): 3-14.

[4] 闫坤如. 人工智能的道德风险及其规避路径[J]. 上海师范大学学报(哲学社会科学版), 2018, 47(2): 40-47.

[5] 陈洪兵. 人工智能刑事主体地位的否定及实践展开——兼评“反智化批判”与“伪批判”之争[J]. 社会科学辑刊, 2021(6): 92-98.

[6] 孙道萃. 人工智能对传统刑法的挑战[N]. 检察日报, 2017-10-22(003).