Open Journal of Legal Science 法学, 2025, 13(11), 2535-2541
Published Online November 2025 in Hans. https://www.hanspub.org/journal/ojls
https://doi.org/10.12677/ojls.2025.1311346

Hans汉斯

强人工智能刑事主体资格的法理批判与归责路径完善

吕泽瑜

济南大学政法学院，山东 济南

收稿日期：2025年10月10日；录用日期：2025年10月20日；发布日期：2025年11月14日

摘　要

本文旨在批判性审视赋予强人工智能刑事主体资格的理论主张，并从法理层面系统论证其不适宜性。依据其自主性、学习能力与决策不可预测性划分为不同层级以构建分析框架。在对学界“肯定说”的核心论据进行了充分地吸纳与细致的辩驳，指出其无论在法哲学基础、法律拟制类比还是风险规制功能上均存在难以克服的缺陷。强人工智能本质上缺乏基于社会实践的主体性与刑法所要求的自由意志及刑事责任能力，对其施加刑罚将导致报应与预防目的双重落空。因此，本文主张应坚持人类中心主义的刑法立场，否定其刑事主体资格，并转而构建一个针对不同层级AI风险特征的、以研发者、生产者、使用者为核心的多层次归责体系，并辅以社会化风险分担机制，以此形成更具现实指导意义的刑事规制路径。

关键词

强人工智能，刑事主体资格，刑事责任能力，刑罚目的，归责路径

Jurisprudential Critique of the Criminal Subject Qualification of Strong Artificial Intelligence and Improvement of Attribution Paths

Zeyu Lyu

School of Political Science and Law, Jinan University, Jinan Shandong

Received: October 10, 2025; accepted: October 20, 2025; published: November 14, 2025

Abstract

This paper aims to critically examine the theoretical propositions for granting criminal subject

文章引用: 吕泽瑜. 强人工智能刑事主体资格的法理批判与归责路径完善[J]. 法学, 2025, 13(11): 2535-2541.
DOI: 10.12677/ojls.2025.1311346

qualification to strong artificial intelligence and to systematically demonstrate its inappropriateness from a jurisprudential perspective. By categorizing strong AI into different levels based on its autonomy, learning capability, and decision-making unpredictability, an analytical framework is constructed. Through fully absorbing and meticulously refuting the core arguments of the "affirmative view" in academia, this paper points out its insurmountable deficiencies in legal-philosophical foundations, analogies of legal fiction, and the functionality of risk regulation. Essentially, strong AI lacks the subjectivity based on social practice, as well as the free will and criminal capacity required by criminal law. Imposing punishment on it leads to the dual ineffectiveness of both retributive and preventive purposes of punishment. Therefore, this paper argues for adhering to an anthropocentric standpoint in criminal law and denying criminal subject qualification to strong AI. Instead, it advocates for constructing a multi-layered attribution system tailored to the risk characteristics of different AI levels, centered on developers, producers, and users, supplemented by socialized risk-sharing mechanisms, thereby forming a more practically instructive criminal regulatory path.

Keywords

Strong Artificial Intelligence, Criminal Subject Qualification, Criminal Capacity, Purposes of Punishment, Attribution Paths

1. 问题的提出

当前，以深度学习、大规模神经网络及类脑智能计算为代表的技术正呈现指数级发展。构建具备自主意识与意志、能够独立进行推理决策并实施复杂行为的强智能体，已逐渐从一个哲学思辨议题转变为具有现实可能性的技术目标。一旦此类强人工智能成为现实，其决策与行为将源于自身的意图与价值判断，从而可能在某些场景下超越人类预设的设计框架与控制范式。

“强人工智能”与仅能模拟特定智能行为、本质上作为工具的“弱人工智能”不同，其核心特征在于其拥有真正的意识、理解力与意向性。这意味着它不仅能处理信息和执行任务，更能理解任务背后的意义，并具有自我意识与主观体验。人工智能在作为关键驱动力推动社会变革的同时，与之伴生的刑事风险已从理论担忧转化为现实问题。这些风险主要呈现为两种形态：一是作为“犯罪工具”，即传统犯罪利用人工智能技术提升其效率与规避能力；二是作为“行为实体”，如自动驾驶汽车自主决策所导致的交通肇事。面对后者，是否追究其刑事责任的前提，是必须解决其刑事主体资格这一核心症结。主体资格是存在者成为实定法主体和具体法律关系的前提[1]。因此，对人工智能法律人格的认定，已成为规制相关刑事风险并构建相应法律框架的逻辑起点。

根据不同标准和角度，有学者根据人工智能体的外部形状将其分为类人型和非类人型人工智能体[2]。根据数据的输入方式分为符号型和数据驱动型[3]。较为普遍的是根据人工智能体的智能水平，将其分为“弱和强”两个等级[4]。强人工智能并非一个均质的概念。根据其自主性水平、学习能力及决策的不可预测性，可以大致划分为不同的层级：从具有高度自主性但仍严格受制于预设目标和约束的人类监督下的人工智能，到具备自我改进和目标更新能力、其行为可能超出开发者预期的自适应人工智能，乃至理论上可能具备完全自主意识和价值体系的“超级智能 AI”。不同层级的强人工智能系统引发的归责难题各异，法律对策也需具备针对性。本文的批判与构建主要针对当前技术发展路径下最可能出现的、具备相当自主性但仍非全知全能的强人工智能形态。

2. 理论争鸣

当前学界关于强人工智能能否成为刑事责任主体，主要形成肯定与否定两种对立学说，其争论深刻反映了不同立场下对技术、法律与伦理关系的不同理解。

2.1. 肯定说：自主行为与规制需求的逻辑证成

随着技术的发展，未来强人工智能系统将具备高度的自主性，能够从外部环境持续获取信息，通过自我学习与迭代，实现认知的深化与行为的创新。肯定论者援引自然人或单位作为先例，以论证强人工智能的主体资格。基于其高度的自主辨认与控制能力，强人工智能在理论上已具备成为刑事责任主体的必要条件。具备自主意识的人工智能体可将犯罪意志作用于现实，转化为犯罪行为，而生命并不是刑事责任主体的必备要件。除生命外自然人和智能机器人在刑事责任能力方面并没有本质的区别[5]。参照单位作为非自然人主体被法律规定为刑事责任主体的法理先例，强人工智能的刑事主体性在逻辑上便不能被全然否定[6]。单位犯罪的理论承认了“组织体”可以具有独立的犯罪意思和犯罪行为为理解强人工智能的“自主意志”提供了类比基础。相应地，其承担刑事责任的方式亦可类比单位的“双罚制”，体现为对系统本身的制裁，例如删除数据、修改程序乃至永久销毁。另一论证路径则基于风险预防与规范填补的现实需求。该观点认为，现有法律框架不足以应对强人工智能引发的刑事风险，在弱人工智能阶段尚具备解释力与实践可行性，但面对具备高度自主性与认知潜能的强人工智能，传统刑罚体系已显现出结构性局限：其既难以对无肉体、无意识的人工智能施加具有伦理意义的刑罚，亦无法妥善回应其行为不可完全归因于特定人类主体的归责困境。因此，从功能性立场出发，承认其法律主体地位便成为一种必要的立法预判与制度安排。有必要进行前瞻性立法，预先构建以强人工智能为规制对象的刑事责任体系，以有效防控潜在的刑事风险。

2.2. 否定说：意志缺失与规范本位的立场重申

在现行法律范式下，人工智能在本质上仍被视为人类意志的延伸与工具，其行为完全源于预设的算法模型与数据驱动，并不具备刑法所要求的独立主观意识与自主意志。否定论认为人工智能只能作为法律客体而不能被视为法律主体，从人类中心主义的基本立场出发，刑事责任的核心在于对具备自由意志的主体行为进行非难，而人工智能因缺乏真正的意志自主，既不具备辨认行为社会意义的意识能力，也不具备基于该辨认而实施或控制行为的意志能力，因而无法成为适格的刑事责任主体。传统刑法理论强调，刑事责任能力是行为人承担刑罚的前提与基础。若对无责任能力者施加刑罚，既无法实现特殊预防的矫正效果，也难以达成一般预防的威慑目的，这与刑罚的正当性基础相悖。因此，在当前以自然人与单位为主体构建的责任体系内，将人工智能直接作为刑事责任主体尚缺乏坚实的法理根基。此外，有学者指出，法学研究应聚焦于解决现实社会问题。在人工智能尚未具备真正自主意识的当下，过度超前地探讨其主体地位并推动相关立法，可能仅是一场缺乏实际规范效用的“象征式立法”，而非回应社会实践需求的理性建构[7]。

3. 核心辩驳：否定强人工智能刑事主体资格的三重维度

在关于强人工智能刑事主体性的争论中，肯定与否定两种立场均与特定时代背景紧密关联。部分持否定观点的学者将讨论植根于当前技术条件与法律框架，而肯定论者则往往着眼于未来强人工智能可能实现的自主形态。

3.1. 对“自主性”事实论断的哲学与法学澄清

肯定论者的逻辑起点在于，未来强人工智能将具备真正的、可与人类相比拟的自主性，能够源于其

自身对环境的复杂理解独立形成犯罪意图并支配行为，传统工具论的解释力将趋于弱化。尤其是对于具备环境学习与适应能力的 AI，其行为已具有一定程度的不可预测性，完全归责于人类主体可能失之公允。高度自主的工具型 AI 和具备环境学习与适应能力的 AI 的出现，确实对以主观过错和因果关系为核心的传统归责模式构成了严峻挑战。然而，这一技术预期需要经过哲学与法学的双重检验。

“自主性”并不等于“主体性”。从哲学层面看，主体性被视为人类区别于其他存在物的根本特征。传统哲学对主体性的理解大多建立在意识哲学或观念论的基础之上，停留在抽象的思辨领域。马克思则从现实的社会实践出发，构建了以自为性、能动性与自主性为核心的主体性分析框架[8]，将主体性奠定在具体、历史的物质实践基础之上。依据马克思的这一实践性框架，人工智能尽管能够模拟智能行为，但其运作从根本上脱离社会实践基础，受制于预设的算法逻辑与数据模型，并不具备基于自身意识与世界互动的自为性，没有在社会关系中形成自我决定的自主性，其“行为”也并非源自内在目的驱动的能动性。因此，人工智能在哲学本质上是人类实践活动的产物与工具，而非真正意义上的主体。

从法学层面看，在人类中心主义的规范立场下，刑法体系的建构与调适须以维护和保障人类根本权益作为最高价值，人工智能产生与发展的根本目的是服务于人类社会的福祉。若在刑法上赋予人工智能以刑事责任主体资格，则在法理逻辑上难以回避对其相应法律权利主体地位的承认。当前肯定论的论证焦点集中于如何令其承担刑事义务与责任，却普遍忽视了权利享有是义务承担不可分割的逻辑前提。这种将义务与权利人为割裂的论证方式，不仅在法理上难以自洽，更在价值层面与人类中心主义的基本原则相悖，导致法律体系在主体制度、归责原理与权利配置上的逻辑混乱。

3.2. 类比单位主体资格的法理局限性

肯定说的另一重要论据来自法律拟制的先例：既然单位可以作为刑事主体，强人工智能同样可以通过法律拟制获得主体资格。单位犯罪的集体意志同样是抽象的、拟制的，这为理解强人工智能的“自主意志”提供了类比基础。承认单位犯罪，本身就突破了刑事责任止于自然人的传统教义，体现了法律适应社会发展的灵活性。这一类比看似有力，实则忽略了单位主体资格的法理根基，尤其刑事主体须具备刑事责任能力这一条件。

不论是自然人主体还是拟制的非自然人主体，其法律地位的确立均非伴随法律诞生而自然存在，而是经由法律制度的明确确认或规范性拟制所建构。传统刑法理论将刑事责任主体严格限定于自然人与单位，其主体资格的赋予始终以具备承担刑事责任的核心能力为根本依据。自然人主体的正当性根植于其自由意志，单位刑事主体资格的正当性根植于其意志与自然人意志之间不可割裂的生成关系。单位作为刑事责任主体，是法律对一种基于自然人集合所形成的“集体单一意志”所进行的有限拟制，其并未脱离以人类意志为原点的责任逻辑，因而也未动摇刑法的人类中心主义根基。

就强人工智能而言，即便其在技术上实现了高度的自主决策能力，即所谓“自由意识”，但其本质与人类的自由意志存在根本差异。强人工智能的“意识”源于研究者在封闭实验环境中，通过集成计算机科学、神经生理学等多学科前沿技术所建构的复杂算法模型。这一过程本质上是对智能的模拟与工程化实现，是其技术复杂性的彰显，而非源于在社会关系中形成的、具备价值反思能力的主体意识。因此，强人工智能即便拥有形式上的辨认与控制能力，此种能力也缺乏内在的社会属性。人类与强人工智能之间并不存在类似于自然人与单位之间那种以意志能力为前提的“代理–被代理”关系。法律上的拟制主体，并非纯粹依靠逻辑推演或形式类比即可成立，而必须立足于具有意志事实的社会实体。因此，单位具备刑事主体性，并不能为强人工智能的刑事主体资格提供法理依据。

3.3. 风险规制路径的功能性质疑

肯定论者从实用主义出发，主张在面对强人工智能引发的、无法归责于特定人类的责任空白时，承

认其主体地位是填补规制漏洞、实现有效预防的必要手段。对于“自适应人工智能”乃至未来可能出现的“具有战略目标设定能力的AI”，传统归责可能面临无人可罚的困境。将AI本身作为责任兜底主体，被视为确保救济、强化预防的务实选择。这一功能主义论证同样存在局限。

首先，刑罚目的的落空。正如贝卡里亚所言：“刑罚的目的既不是要摧残折磨一个感知者，也不是要消除业已犯下的罪行。刑罚的目的仅仅在于：阻止罪犯再重新侵害公民，并规诫其他人不要重蹈覆辙。”[9]当刑罚对象转向强人工智能时，报应与预防都难以有效实现。从报应目的来看，刑罚的施加需使受刑主体感受到痛苦，并以此彰显刑法的严厉性与不可违背性。删除数据、修改程序、永久销毁等措施本质上无法使强人工智能产生痛苦感知，其效果仅相当于对所有者或使用者财产权的限制或剥夺。刑罚权作为国家专属的权力，具有严格的公法属性，若将删除数据、修改程序等列为刑罚措施，无异于允许私人或企业行使本应属于国家的刑罚权。就预防目的而言，刑罚亦难以在强人工智能语境下发挥应有功能。在特殊预防层面，费尔巴哈所提出的“心理强制说”建立在犯罪主体对痛苦具有感知与回避能力的基础上[10]，而强人工智能既无肉体亦无心理感受，无法基于“快乐–痛苦”的权衡机制抑制自身行为。在一般预防层面，刑罚的威慑效果依赖于潜在犯罪主体对刑罚后果的认知与畏惧。若以惩罚人工智能来威慑人类，显然缺乏直接的心理联结；而若以惩罚某一人工智能来威慑其他人工智能，则又因后者不具备情感与痛苦感知能力，难以形成有效的行为遏制。

其次，肯定论的方案可能导致责任规避的反效果。人类作为技术的研发者、受益者与风险最终承担者，应是责任链条的终点，将责任转移给一个无财产、无权利的人工智能，最终可能导致受害者求偿无门，削弱了法律对风险源头的规制力度。人工智能技术的指数级发展，常被视为支持赋予其法律主体资格的“肯定论”之逻辑起点。然而，库兹韦尔在其“奇点理论”中主张，当强人工智能超越人类智能之时，其会代替人类统治世界，人类会在各方面丧失中心地位，该时刻被称为“奇点时刻”[11]，恰恰是这种超越人类控制预期的技术爆炸潜力，构成了反对赋予强人工智能刑事主体资格的核心理据。当前，在弱人工智能的辅助下，人类在诸多领域仍面临无法探知与解决的复杂问题。若强人工智能的智能水平全面超越人类，其认知与行为模式将超出人类的理解范畴与有效监管边界，以人类理性为基础构建的传统法律体系的威慑机制与规范效力将面临失效风险。对人类无法理解、无法预测且无法控制的智能体施加刑罚，非但无法实现刑罚的预防与报应功能，反而会架空刑事责任制度本身，导致刑法的规范权威被消解。因此，从风险预防与刑法效力的角度审视，技术发展的不确定性非但不能证成其主体资格，反而警示我们必须审慎划定其法律边界。

4. 归责路径：风险分配视角下的责任溯源与体系完善

在否定强人工智能刑事主体资格之后，关键在于构建切实有效的归责路径，可行的替代路径并非盲目扩张刑事主体范围，而应在既有法律框架内进行调适与完善。鉴于强人工智能存在不同层级，归责路径亦应体现差异化思路。

4.1. 归责体系的基本原则：容许风险与责任明确

若出现纠纷时将人工智能拟制为责任主体，实则是一种责任转移，这将动摇刑法中罪责自负与行为与责任同时存在的基本原则。无论针对何种层级的强人工智能，归责体系的构建都应遵循以下基本原则。首先，坚守法所容许的风险边界，刑法的介入应保持谦抑。“刑法不能直接将科学技术中具有风险的探索活动予以禁止，也不能在科学技术所带来的风险实现后，追究相关人员的刑事责任”[12]，对于研发者而言，若其技术活动受正当目的支配，所涉风险属于社会发展和科技创新所必须承担的容许风险，则原则上不应予以刑事苛责。其次，确保责任链条的清晰与闭合，防止因人工智能的自主性导致责任缝隙。

通过明确研发、生产、销售、使用各环节主体的法律义务，形成无缝衔接的责任网络，使得任何由其引发的刑事危害都能找到最终的责任承担者，即人或者由人组成的单位。

4.2. 针对“高度自主的工具型人工智能”的归责：强化监督者的责任

此类系统的核心难题源于人机协同的责任界面模糊，一是监督者的功能性懈怠与注意义务虚化。使用者可能因系统常态化可靠运行而产生过度信赖，从而在系统发出接管请求或出现异常状态的关键时刻未能及时干预。例如，在自动驾驶中，驾驶员因长期脱离驾驶任务而导致的情境意识下降与反应迟缓，其过失责任应如何区别于传统驾驶情境，成为司法认定的新难题。二是人机控制权切换中的因果断裂。当法益侵害发生于控制权从系统向人类操作员移交的瞬间，如何精准判定是系统交互设计缺陷(如预警不充分、界面不友好)，还是使用者的个体反应失误构成了结果发生的决定性原因？此种情形极易导致刑事责任链条中因果关系的中断与归属不明。

对于目标和行为边界相对清晰的受控自主的强人工智能，其行为后果在很大程度上可追溯至研发者、生产者或使用者。第一，明确使用者的监督过失责任。应通过立法或司法解释，确立使用者在关键应用场景下的动态注意义务与即时有效接管义务。在发生重大法益侵害时，可考虑适用过错推定原则，由使用者承担其已履行合理监督职责的举证责任。第二，强化研发者与生产者的产品责任。应要求其在人机交互设计中履行高度审慎义务，若事故源于交互界面的设计缺陷或预警机制的失灵，应依据刑法中有关生产、销售不符合安全标准产品罪等条款，追究其产品责任。

4.3. 针对“自适应人工智能”的归责：引入组织责任与风险分担机制

当人工智能具备自我优化和目标更新能力时，传统归责模式面临严峻挑战。行为不可预测性与预见可能性的瓦解，研发者与生产者可以“其行为已超出我们初始设计的预期”为由进行抗辩，主张损害结果不具有主观上的预见可能性，从而规避过失责任。决策过程的“黑箱化”使得其决策逻辑可能复杂到连创造者都无法完全解释，当无法追溯导致危害的具体算法或数据原因时，传统的“行为–责任”因果链条便难以建立。在无法归责于人工智能本身，又难以追究具体研发人员或个人用户刑事责任的情况下，就会出现无人承担刑事责任的空白地带。

对于行为可能超出开发者明确预期的强人工智能，归责难度增加，需引入更灵活的机制。其一，探索组织体刑事责任。当难以将危害结果直接归因于某个具体的研发人员或用户的过失时，可以借鉴单位犯罪的原理，追究研发企业或运营企业作为组织体的刑事责任。前提是该企业未能建立有效的伦理治理框架、风险评估机制和事后干预措施，存在整体性的管理失职，从而制造了法所不容许的严重风险。其二，构建社会化的风险分担体系。为应对无法清晰归责于特定人类主体的意外风险，尤其是强人工智能在迭代中产生的不可预见的危害，必须建立强制性的保险制度或行业性赔偿基金。这可以作为刑事归责体系的必要补充，确保受害人能够及时获得救济，同时分散技术创新带来的社会成本。

5. 结语

强人工智能刑事主体资格之辩折射出法律面对技术革命时的根本立场选择。本文从刑事主体性理论、刑事责任能力以及刑罚目的角度否定强人工智能刑事主体性：从主体性看，其缺乏在社会关系中形成的价值认知与伦理判断；从责任能力看，算法决策与刑法要求的自由意志具有本质区别；从刑罚目的看，对其施加制裁既无报应意义，也难达预防效果。刑事责任的边界必须止于人类主体。面对强人工智能带来的新型风险，创造电子主体并非权宜之计，而应回归责任本源，构建精准化的归责体系。这一体系既包括对研发者、生产者、使用者等人类主体义务的明确界定，也需引入适应技术特性的风险分散机制。未来法律的发展，应当在保持人类中心主义范式的前提下，通过责任规则的创新与细化，实现技术赋能

与风险管控的平衡。

参考文献

[1] 陈劲松. 传统法律主体资格标准理论及其当代变革[J]. 学术交流, 2019(7): 84-96.

[2] 刘宪权. 对强智能机器人刑事责任主体地位否定说的回应[J]. 法学评论, 2019, 37(5): 113-121.

[3] [英]亨利·布莱顿, 霍华德·塞林那. 视读人工智能[M]. 张锦, 译. 合肥: 安徽文艺出版社, 2007: 3.

[4] 王耀彬. 类人型人工智能实体的刑事责任主体资格审视[J]. 西安交通大学学报(社会科学版), 2019, 39(1): 138-144.

[5] 魏东. 人工智能犯罪的可归责主体探究[J]. 理论探索, 2019(5): 5-13.

[6] 刘宪权. 关于人工智能时代刑事责任主体演变理论研究的再辨析[J]. 法学, 2025(9): 82-94.

[7] 马荣春. 中国刑法学研究主体性的实现路径——由人工智能犯罪主体化问题再出发[J]. 关东学刊, 2025(1): 20-39.

[8] 秦铭. 生成式人工智能的“主体性”批判——基于马克思主体性思想的哲学视阈[J]. 中国地质大学学报(社会科学版), 2025(8): 1-9.

[9] [意]切萨雷·贝卡里亚. 论犯罪与刑罚[M]. 黄风, 译. 北京: 商务印书馆, 2017: 49.

[10] 徐久生. 费尔巴哈的刑法思想——费氏眼中的刑法与社会[J]. 北方法学, 2013, 7(5): 91-100.

[11] 朱彦明. 奇点理论: 技术“复魅”世界?——批判地阅读库兹韦尔的《奇点临近》[J]. 科学技术哲学研究, 2020, 37(6): 83-88.

[12] 陈兴良. 风险刑法理论的法教义学批判[J]. 中外法学, 2014, 26(1): 103-127.