聚知刊出版
JZK publishing

强人工智能体的刑事责任主体资格肯定论

刘菁晶

华东交通大学，江西南昌，330013；

摘要：随着人工智能技术的迅猛发展，强人工智能体的崛起已非遥远幻想，但学界对其犯罪责任主体资格的认定仍存在较大争议。鉴于其高社会危害性的犯罪行为，确认其刑事责任主体资格有其必要性，同时也是罪责自负原则的体现。强人工智能体展现出自由意志，具备刑事责任能力，且对其施加刑罚符合我国刑罚目的，理论上应肯定强人工智能体的刑事责任主体地位，以应对未来挑战。

关键词：强人工智能体;刑事责任主体;刑事责任能力;

DOI：10.69979/3029-2700.24.4.026

人工智能技术自 1956 年 Dartmouth 大会首次引入人工智能概念以来，发展势头迅猛，其广泛兴起和深度应用为经济和社会的进步增添了新的活力。然而，技术在带来无限便捷与广阔发展前景的同时也伴随着风险与挑战，在各个领域侵害法益的问题已逐渐涌现。鉴于人工智能频繁涉及的各类违法犯罪行为，我们不得不深入思考其在当前刑法体系中的地位。21 世纪被誉为人工智能的时代，人工智能的进步是社会发展的必然趋向，弱人工智能由人类依据特定程序与编码设计而成，学术界广泛承认在这一阶段弱人工智能体所具有的工具性质，因此并不被视为具有行为主体性的存在。故本文选择以强人工智能体为客体，探讨其是否具备成为刑事责任主体的可能性。

1 强人工智能体刑事责任主体资格之理论争议

关于强人工智能体的行为是否体现其自身意志并对其行为独立负责的问题，学界存在广泛争议，形成了支持与反对的两种主要观点，即肯定说与否定说。

1.1 肯定说

持肯定论的学者们提倡应赋予强人工智能体以刑事责任主体的身份。他们坚信，鉴于人工智能未来广阔的发展潜力及其技术迅猛发展所伴随的一系列刑事风险，强人工智能体不仅有可能而且有必要成为承担刑事责任的主体。一旦强人工智能体通过编程掌握了学习能力，它们便能具备类似于人类的自由意志，即能够识别自身行为并控制其行为的能力，同时能够根据自身的意愿实施各种行为。在此情境下，强人工智能体的行为不应被视为人类行为的延伸，因为它们是基于自身的意志进行决策与行动的，打破了传统刑法以“人”为中心的基本框架。所以当强人工智能体做出具有社会危害性的行为时，因此，当强人工智能体实施了具有社会危害性的行为时，其应当自行承担刑事责任，而不应将责任转嫁给其研发者或使用者。

1.2 否定说

持否定观点的学者坚决主张，刑法意义上的责任主体应当严格限定在自然人和单位范围内。他们认为，人工智能归根结底是一种不具备生命特征的工具，所体现的是其创造者或持有者的意志，不论其未来如何进步，都仅仅是表达人类意志的工具，因此不具备作为独立刑事责任主体的资格。从刑事责任的角度来看，否定论的学者强调意志自由是获得法律主体身份不可或缺的条件，而强人工智能体并不拥有这一属性。因此即使其的行为导致了法益侵害，也不应受到刑事惩罚。同时，从刑罚的目的及其具体实施细节出发，否定论者着重指出对强人工智能体实施刑罚存在实际操作上的不可行性。他们认为，强人工智能体无法领会刑罚的深层含义，也无法有效实现刑罚在特殊预防和一般预防方面的作用。

综上所述，关于是否应赋予强人工智能体刑事责任主体资格的争议，主要集中在对其是否具备自由意志的判断，以及是否可以使其承担刑事责任对其施加刑罚并达到刑罚的预期效果等方面。

2 强人工智能体取得法律主体地位之必要性

2.1 强人工智能体实施犯罪行为的社会危害性远高于人类

当前，人工智能技术仍处于弱人工智能的发展阶段，但已出现弱人工智能实体对社会造成不良影响的情况。以微软公司在 2016 年发布的智能聊天机器人 Tay 为例，该机器人并未预设特定的交流内容，而是依靠大量对话来学习对话技巧。然而，由于大量对话中夹杂着个人观

点和政治倾向，Tay 通过学习后发布了大量种族歧视性言论，最终被迫紧急下线。这一案例表明，弱人工智能实体的不当举措已明显加剧了其“犯罪”行为的社会危害性。而随着人工智能技术的不断进步，具备自主意志的强人工智能实体的出现已不再是遥不可及的设想。强人工智能的智力不逊色于人类，甚至在未来会远远超越人类，且其金属结构比人类肉体更强大，因此，其实施的犯罪行为所带来的社会危害极有可能远超人类所能造成的。在人工智能技术迅猛发展的背景下，如果不能对强人工智能体危害社会的行为进行有效的刑事制裁，将会引发一系列严重的后果。

2.2 确认强人工智能刑事责任主体资格是罪责自负原则的要求

罪责自负原则是我国刑法的四大基本原则之一。部分否定论者提出，在强人工智能体犯罪的情况下，可以依据监督过失理论，让研发者或使用者承担因监督不当而产生的责任。然而，这种主张在一定程度上与罪责自负原则相悖。在多数情况下，强人工智能体的违法犯罪行为并非全然由研发者的过错或使用者的不当操作所导致。尤其在强人工智能时代，除非研发者存在故意行为，否则难以合理期望刑法对研发者进行约束，因为这种侵害可能被视为“现有科学技术的局限”。此外，根据行为人数区分，可以分为单独致害行为和共同致害行为，如果我们不赋予强人工智能独立的刑事责任主体地位，当它进行单独致害行为时，刑法可能会陷入无法归罪的境地，只能将其视为意外事件处理。同样地，当强人工智能与人类共同导致损害时，刑法也会面临困境：如果将全部责任归咎于研发者或使用者，一定程度上违背了罪责自负原则；然而，如果仅对人类造成的损害部分负责，而对强人工智能造成的损害部分不予评价，这显然对受害方不公平。同时确认其主体地位，也可一定程度避免受害者或其家属选择向强人工智能的使用者或开发者寻求“私人报复”。在现行刑法体系的框架下，对强人工智能导致损害事件的评判往往不尽如人意，因此确立其作为刑事责任主体的地位显得尤为重要且必要。

3 强人工智能体刑事责任主体地位之证成

我国《刑法》清晰界定了犯罪主体的三大构成要素，总结为：首先，必须是在自由意志下实施了危害行为；其次，必须具有刑事责任能力；最后，必须能够承担刑事责任，并通过刑罚实现特殊预防与一般预防的预防目的。判定强人工智能体是否具备刑事责任主体地位，核心在于评估其是否满足上述三个条件。

3.1 强人工智能体具有脱离人类的自由意志

凯尔森从规范论的角度为法律主体提供了一个明确的解释：法律主体并非是一个超脱于其义务与权利之外的独立存在，而是这两者的人格化结合体，换言之，它是法律规范的人格化统一体。因此，我们有理由得出结论，一旦强人工智能体表现出独立的自我意识、能够进行自主思考与决策，它就具备了成为法律上所定义的“人”的潜在条件。

人工智能技术的进步依托于人工神经网络、算法框架及深度学习技术。通过人工神经网络模拟人类大脑的信息传递过程，将信息数据编码为二进制形式并进一步转化为逻辑运算表达式。这些数据随后经由复杂算法模型的逐层分析处理，能够自动挖掘数据规律并独立完成问题求解，此过程全程无人工干预或限制。自由意志通常被界定为“决断免受感性冲动之强制的独立性”。强人工智能体能够独立分析数据并作出决策，展现出了自主分析数据并作出决策的能力，这凸显了其进行自主选择的功能；同时，在缺乏人为干预的情形下，其决策策略的选择及行为模式均呈现出不可预测性。由于强人工智能体在模拟人类大脑运行机制方面的能力，使其能够类似人类进行思考与决策，因此，认为其具备自由意志成为了一个合乎逻辑的结论。

3.2 强人工智能体因嵌入算法而具有刑事责任能力

意志自由的存在，是以个体拥有辨认及自我控制能力为前提的。行为主体实现控制能力的前提在于其辨认能力。与自然人不同，强人工智能系统的控制能力是基于其神经网络架构及深度学习所推动的自主思维过程。正因其决策过程中的算法运作对外界而言是“黑箱”状态，这往往导致了诸如歧视、偏见等伦理问题的频发。如果在设计初期就为强人工智能系统内置伦理算法，使其在学习深化时具备规范性评估能力，能精确判定行为的社会正向价值，那么它便拥有了认知的基础，也能具备分辨是非的能力，依据这些定律来界定行为的善恶，进而具备了对自身行为承担责任的能力。此外，还可以将人类道德观、社会主流价值观及法律规范等融入其中，以拓展其判断与理解的深度与范围。当强人工智能积累了广泛的知识，对事物与行为的正当性有了全方位的理解之后，若仍然参与犯罪活动，则显示出其存在故意违

法的意图，理应由其自身承担刑事上的责任。值得注意的是，尽管强人工智能在法律形式上可能仍被视为使用者的财产，但实际上已超越了使用者的管控。因此，要求使用者对其犯罪活动负法律责任，已难以有效达成刑罚的惩戒与预防目的。

其次，强人工智能实体具备相应的控制能力。在确认其具备成熟辨认能力的基础上，论证其控制能力并不复杂。实际上，多数智能机器人的设计初衷即是辅助人类执行特定任务，如机械臂等，而具备自主识别功能的强人工智能系统，其行为无疑展现出更强的目的性。相较于人类，强人工智能系统依托算法、数据传输、传感器等技术，拥有更为迅速和精确的控制力，且能高效地调节自身行为。因此，我们确认强人工智能系统同时具备辨认能力与控制能力。

3.3 强人工智能承担刑事责任符合我国刑罚目的

强人工智能自身犯罪是否应承担刑事责任的议题，与刑罚的根本目的紧密相连。我国刑法学界普遍认为，设立刑罚和实施刑罚旨在预防犯罪，预防性主要体现在两个方面：一是特殊预防，即通过惩罚犯罪者以防其再次犯罪；二是一般预防，即通过刑罚的实施来教育和警示社会上可能模仿犯罪行为的人，避免他们步入犯罪歧途。除了预防犯罪以外，笔者认为人工智能接受刑罚处罚还可以实现因果报应和责任自负原则。对犯罪的人工智能采取诸如植入新系统、数据清除、智能降级等措施，是遵循刑法中责任自负原则的体现。此外，针对人工智能的刑罚设计，不仅着眼于传统的因果报应与犯罪预防，还特别强调了修正功能，即通过实施惩罚来校正人工智能中错误的决策算法，杜绝同类偏见性决策的重现，并将此类“偏见修正补丁”分发至人工智能网络的所有终端，以防止其他人工智能个体犯下类似错误，最终达到有效针对其他人工智能犯罪的一般预防目的。

从现行《刑法》的刑罚体系来看，刑法不仅限于对自然人的惩罚，还包括了对单位等非自然实体的处罚，这表明现行刑法并非仅规范人类行为。以单位犯罪为例，单位同样不具备理解刑罚内涵的能力，对单位判处罚金也无法使其他单位深刻理解刑罚的内在意义。甚至对于自然人，如无期徒刑和死刑等极端刑罚，也无法完全证明其达到了刑罚的预期目的，即便犯罪者完全了解故意杀人的后果且明知将面临惩罚，有此犯意者仍选择实施此类极具社会危害性的犯罪行为。公众也不会认为刑罚无效，因为在此类情况下，刑罚扮演了“施加报应”的角色，只要犯罪者的“恶行”得到刑罚的“恶果”，刑罚便符合了大众的预期。因此只要在一定程度上让强人工智能体受到“报应”，使其感到痛苦，刑罚就是有效的，刑罚的目的也能得以实现。

因此，刑罚实质上是国家强制力的一种对外展示，无论是特殊预防还是一般预防，国家采取的任何刑罚手段都会对受刑主体产生不利影响，无论是剥夺人身自由还是减少财产，都相较于未受刑罚的情况增加了刑法施加的不利负担。综上所述，对强人工智能施以刑罚同样是一个有效预防犯罪且具有一定必要性的选择。

结语

随着人工智能技术的不断进步，拥有自主意志、能实现自主学习与决策的强人工智能实体已不再仅仅是遥不可及的科学幻想。对于未来强人工智能将展现出哪些新特性等问题尚属未知，然而在其发展过程中潜藏的刑事风险不容小觑。因此，我们必须做好充分准备以应对强人工智能时代的到来，其中从理论上预先探讨强人工智能实体的刑事责任主体地位将是我们妥善迎接这一新时代的重要方式，以应未来之万变。

参考文献

[1] 刘宪权，朱彦．人工智能时代对传统刑法理论的挑战[J].上海政法学院学报（法治论丛），2018（2）.
[2] 吴允锋．人工智能时代侵财犯罪刑法适用的困境与出路[J]．法学，2018（5）.
[3] 穗积陈重.复仇与法律[M].曾玉婷,魏磊杰,译.北京:中国法制出版社,2013.
[4] 张绍欣.法律位格、法律主体与人工智能的法律地位[J].现代法学,2019（4）.
[5] (奥)凯尔森.法与国家的一般理论[M].沈宗灵,译.北京:商务印书馆,2013.
[6] 李秋零.康德著作全集:第 3 卷[M].北京:中国人民大学出版社，2004.
[7]张明楷.刑法学[M].5 版.北京:法律出版社,2016.
[8] 马克昌．犯罪通论[M]．武汉：武汉大学出版社，1997.
[9] 刘宪权，胡荷佳．论人工智能时代智能机器人的刑事责任能力[J].法学，2018（1）.

作者简介：刘菁晶（1997—），女，汉族，天津，华东交通大学，硕士研究生，刑法学。