技术逻辑下强智能机器人刑事责任的再审视*

赵天水

（天津财经大学法学院，天津　300222）

[摘　要] 目前学界多从法律逻辑角度分析强智能机器人的刑事责任，鲜见从技术逻辑切入研究的成果。强智能机器人能否出现是讨论其刑事责任的前提，也是否定论对其加以批判的主要原因。人工智能技术的指数级增长与涌现使得强智能机器人的出现具有内动力，只有内动力而无需求亦很难将“梦幻”变为“现实”。在军事竞赛、商业化对强智能技术升级存在渴求的政治环境和市场环境下，强智能机器人的出现便具有了外动力。在技术内、外动力拉动下，强智能机器人的出现只是一个时间问题。现有对强智能机器人刑事责任的质疑，本质上是人类中心主义滑向了形而上学的机械、静止观，使得刑事理论研究无法为未来的刑事立法提供智识补给。在重构强智能机器人刑事责任过程中，亟需解决现有研究中主体内涵阙如、刑事责任认定标准不当和权利保留范围过宽的问题。根据强智能机器人硬件、软件与湿件的特点设置差异性的剥夺感可达到区分制裁梯度的效果，对强智能机器人犯罪的罪名可比照自然人犯罪进行。不过基于处罚成本的考虑，应以对硬件、软件的处罚为原则，以对湿件的处罚为例外，在各种处罚内部又应坚持以轻缓处罚为原则、以严厉处罚为例外。规制强智能机器人的前置法宜采取单独立法模式，在前置法中对智能机器人概念、主体种类、技术更新层次及标准、违法行为表现、处罚种类和等级进行规定，突出对人工智能技术标准的描述和专业技术人员鉴定程序的规范化。

[关键词] 强智能机器人　技术逻辑　涌现　人类中心主义　重构

[中图分类号] D924；TP18　**[文献标识码]** A　**[文章编号]** 2096-983X（2023）05-0094-09

21世纪是人工智能时代，人工智能产品早已渗入衣食住行的方方面面，人类生活与人工智能实现了高度融合。对于人工智能的发展趋势，最著名的当属库兹韦尔预测到2045年会出现技术奇点，彼时地球上非生物智能的数量将远远超过人口总量，此“惊世骇俗”观点遭到诸多批判。①对于我国人工智能发展现状及强人工智能的未来，中国科学院院士、清华大学人工智能研究院院长张钹表示，“据统计，全球10亿美金估值以上的人工智能企业一共有40家，中国有15家。就人工智能的发展来看，我国和国际上的发展差距并不是很大。将来机器人会不会成为我们的主人？现在，大家越来越认识到这个可能性是存在的，阿西莫夫、霍金也

收稿日期：2022-09-08；**修回日期：**2023-05-31

***基金项目：**2021年度最高人民检察院检察理论研究课题“涉外网络犯罪刑事管辖权研究”（GJ2021D27）；2021年度天津市教委科研计划项目“人工智能时代侵财犯罪刑法适用的困境与出路”（2021SK111）

作者简介：赵天水，刑法学博士，副教授，硕士研究生导师，主要从事科技犯罪研究。

①本文认同技术奇点必然会到来，但对到来的时间持不同看法。强智能机器人的出现是个时间早晚的问题，不见得一定是2045年。

提出过相关观点，如果机器的智能超过我们，我们就失去对它控制，对于这样的机器人要加以限制，加以治理。”[1]中国工程院院士王天然认为，“人机共融是将来机器人技术发展的重要方面。未来，人与机器将‘宛如同类’，是最好的伙伴关系。”中国工程院院士李培根也表示，“与第一代和第二代的智能机器人不同，第三代机器人将以智能协作为特点，可通过数据分析、认知学习、人机交互、自然语音处理等与人实现情感互动。而第四代机器人将更进一步，通过持续学习、协同学习、场景自适应等，实现‘自主服务’。”[2]2023年3月15日，OpenAI宣布推出ChatGPT4，OpenAI于同月27日发布了GPT-4的全面简报，其中包括一组考试结果，GPT-4模型可以超越90%考生的水平，轻轻松松便能考进哈佛、斯坦福等常春藤名校。与诸多软件的智能水平仅停留在单一领域不同，GPT-4不仅能够处理文本信息，还能快速高效地识别图像、视频并作出进一步反应。不能否认，GPT-4依然在社会偏见、幻觉和对抗性提示等方面还存在局限性，但其所体现出的智能水平意味着我们所处时代的科技水平离奇点越来越近。[3]概言之，在人工智能逐步从第一、二代向第三、四代发展的过程中，强人工智能的出现是存在可能性的，对其刑事责任进行讨论便具有了现实意义。以对强智能机器人刑事责任的讨论为例，持否定意见者（简称否定论）首先就对将来能否出现强智能机器人表示怀疑，认为该命题是个伪命题。众所周知，无科技则无人工智能，对强智能机器人刑事责任的讨论不仅关乎法律逻辑层面的分析，更关乎技术逻辑方面的解构。然而，现有对强智能机器人刑事责任的讨论多从法律逻辑角度切入，鲜见从技术逻辑角度进行的分析，这不能不说是一种缺憾。基于此，本文尝试从技术逻辑出发与否定论进行商榷，呼吁刑法学界正视人工智能的技术特点，不能以未来未到来为由去封闭强智能机器人刑事责任的讨论空间。本文仅为抛砖引玉之作，不足之处还望求教于方家。

一、技术逻辑下强智能机器人的出现并非梦幻

强智能机器人包含类人类智能机器人和超人类智能机器人。对于此类机器人能否在未来出现，否定论者可谓是技术悲观主义者，他们始终认为智能机器人仅是人类的工具产品，不会产生自主意识，无法具备人类的智能水平，对强智能机器人刑事责任进行研究是在制造学术泡沫，严重偏离了刑法教义学划定的解释轨道，是刑法研究的倒退。[4](P119)如果基于现有人工智能技术水平去观察，智能机器人确实还未出现自主意识，否定论的上述观点不无道理。但需要注意的是，强智能机器人的出现并非研究者主观臆想的产物，而是存在坚实的技术逻辑。

（一）技术内动力：指数级增长、“涌现”的双重催化

第一、二次工业革命以蒸汽机、电力为标志，技术呈现出直线性增长特点。与前两次工业革命不同，以信息技术、生物技术等为内容的第三次工业革命在前期可能增长极为缓慢，但在克服技术瓶颈后则会以难以想象的速度增长，这种技术特点被称为指数级增长。指数级增长源于极微小的增长，前期很难有人留意到它的增长幅度和它本身，随后会不经意间呈现出爆炸式地增长而让人完全出乎意料。[5](P1)如生物学家花整整一年时间仅破译人类基因组的万分之一，按照这个速度最少得一百年才能完成，但在解决关键技术问题后仅花费十五年便完成任务。同理，虽然目前人工智能技术在制造机器人的自主意识方面还存在技术瓶颈，但不能忽视技术瓶颈正是将来出现强智能机器人的注脚，此技术瓶颈克服后便会出现指数级的增长态势，人工智能技术大爆炸难以避免。

强智能机器人能否出现的关键在于如何制造出自主意识。笛卡尔将精神从肉体中剥离出来建立二元论，在人的肉体之外能否出现自主意识？这无疑成为技术悲观主义者不看好强智能机器人的关键原因。如有否定论者认为智能

机器人运行的基础是人类编写的算法，即便将来智能机器人产生刑事风险，那风险也是源于算法偏差而非机器人的意识控制。[6](P143-144)该观点暴露出论者对人工智能领域如何实现自主意识的途径不熟悉，现有实现智能机器人自主意识的途径有三大学派：符号学派、连接学派和行为学派。符号学派模拟智能软件，连接学派模拟大脑硬件，行为学派模拟人的身体，三者分别是从功能、结构、身体的角度来实现自主意识。符号学派假设自主意识是先验地存储于算法黑箱之中，是通过自上而下的设计实现，现有技术已经证明这条道路难以出现自主意识。后两种学派都坚信自主意识是通过自下而上涌现出来的。[7](P18-22)涌现指的是在人工智能技术发展过程中产生的新特性，而此新特性并不包含在系统组成部分或先前功能状态中。如单个的蛋白质分子不具备生命特征，但是大量的蛋白质分子组合在一起形成细胞的时候，整个系统就具备了“活”性，这就是典型的涌现。概言之，在人工智能技术发展过程中，不经意间会出现难以预见的强智能状态。以全脑仿真为例进行说明，全脑仿真就是用非生物基质仿照某个大脑制造一个同等运行功能的副本，其以大脑的可塑性作为生物学基础，在人类成长过程中婴儿和胎儿的大脑结构、神经元的轴突和突触功能都会出现重构。仿照人脑结构，遵循测绘、模拟和实体化步骤，通过运用遗传算法、仿生算法、克林尼递归定理等自省程序可写出具有自省、自我觉知能力的程序，正是用这些算法程序重建人脑的复杂性使智能从这些基本的计算单元的交互中涌现出来。当智能机器人产生自我指涉的涌现时，智能机器人便具有了自主意识。[8](P172-175)

（二）技术外动力：军事竞赛、商业化对强智能技术的渴求

在目前地缘政治背景下，各国间的技术竞争不仅关乎科技创新程度，更关乎国家安全。任何一个国家都难以阻止其他国家去制造用于战争的强智能机器人，这并非猜想，毕竟强智能机器人比士兵更准确、更少犯错，可减少所谓的附带损失。智能机器人以算法为核心，在制造者赋予其自我意识并将杀戮规则及对己方的忠诚写入算法时，无疑会吹响军事竞赛的号角。将来随着强智能机器人在军事领域的技术日益成熟，将技术应用于日常生活的需求便会被制造出来，从而产生强智能机器人的商业化。不过，不用过于担心军事竞赛会使强智能机器人变成杀戮机器进而毁灭人类，智能机器人的“意识”源于算法，回报函数会使算法的计算结果朝着设计目标最大化发展，“如果人工智能通过强大的优化来实现，并允许其通过反复自我改善增强智能，那么它的行为就不会受到人类行为的引导。它作出的每个行为，提供的每条建议，其核心都是坚决地实现回报函数最大化。”[9](P207)只要回报函数中含有善良伦理的算法规则，生活中的强智能机器人即便产生自主意识，亦会遵循起初的算法规则。

有否定论者认为，强智能机器人承担刑事责任的前提为“不可控制”，与设置“删除数据、修改程序”的刑罚措施来使其可控制之间存在技术逻辑层面的矛盾。[10](P94-95)在本文看来，这其实是对算法、回报函数等技术规则的误读，人类虽然可以将善良等伦理规则写入算法并增大回报函数，但同时不能忽视上述谈及的涌现有使智能机器人产生自主意识的技术空间。换言之，从算法写入、回报函数输入方面，人类是人工智能的设计者，人类对智能机器人拥有控制力；但在智能机器人产生自主意识后，基于让其能按照人的社会规则运行，便需要运用技术手段让其再度可控。因而，二者之间并不存在技术逻辑上的矛盾，不会对强智能机器人的商业化构成障碍。

在本文看来，人工智能技术的指数级增长与涌现使得强智能机器人的出现具有内动力，只有内动力而无需求亦很难将“梦幻”变为“现实”。在军事竞赛、商业化对强智能技术升级存在渴求的政治环境和市场环境下，强智能机器人的出现便具有了外动力。在技术内、外动力拉

动下，强智能机器人的出现只是一个时间问题。众所周知，法律需要对社会治理需求进行回应，刑法亦不例外，强人工智能机器人的出现必然会颠覆现有以自然人的辨认能力、控制能力为核心概念而建构起来的刑事责任体系（后文详述，此处不再赘述）。因而，在奇点未来之前对强智能机器人的刑事责任进行研究，便具有了刑法研究的实践价值。

二、技术逻辑下强智能机器人刑事责任面临的质疑及回应

虽然有学者已经认识到对强智能机器人的刑事责任进行分析时需要区分技术逻辑与法律逻辑，[11](P115)但在论证过程中依然没有跳出以人类智能作为标准去衡量人工智能的传统路径，在将强智能技术描述为神秘主义和不可知论的同时却偏向了否定论的阵营。应当说，此类观点在否定论中具有代表性，为了澄清质疑、正本清源，笔者对否定论提出的质疑进行梳理并予以回应。在观察中不难发现，无论否定论提出何种反对意见，本质上都是在拿“现有”“人”的智能特点和刑事司法特点作为衡量强智能机器人刑事责任的尺度。

（一）面临的质疑

1. 质疑一：强智能机器人无法具备“人”的刑事责任要素

观点一：智能存在本质差异，无法感知刑罚恐惧，无预防效果。论者以人类存在语言认知、思维认知、文化认知等深层次的神经认知且这些认知能够让人感知痛苦，人工智能所具有的均是浅层次认知进而无法具有神经认知并感受痛苦。刑罚的功能是惩罚与预防，发挥作用的前提是能够发挥威慑作用，以能感受到刑罚痛苦和内心恐惧为基础，因而对强智能机器人难以实现刑罚功能。[12](P105-108)

观点二：智能机器人无辨认能力与控制能力。在辨认能力方面，论者以人拥有理性能力和语言能力为由认为强智能机器人难以获得这类人类专属能力，而缺乏这类能力会导致难以取得辨认能力。[13](P200)在控制能力方面，与人工智能单一集中在道路交通等某些特定领域不同，人的控制能力范围具有开放性；人工智能会随着算法规则的不同产生不同结果，使得对控制能力判断具有了恣意性；人的意志具有理性因素，更重要的是还有非理性认知，而人工智能的判断依靠的是纯粹的数理计算。[14](P67)

否定论上述观点无一不在揭示其人类中心主义立场，在他们看来，“人类中心主义的责任体系具有恒久适应力，不能因应前沿科技而将刑法重构为技术管理法，更不能将之建立在修辞和想象之上。”[15](P7-9)将人作为万物的尺度，是人主宰世界的体现，将其体现在立法中具有合理性。然而，在分析强智能机器人是否应有刑事责任时，不能混淆核心问题。

依据否定论，强智能机器人难以拥有刑事责任的根本原因在于其没有人类的智能水平且没有辨认能力和控制能力，实际上否定论者还是在论证强智能机器人不是“人”，却对强智能机器人为何不应具有刑事责任未能展开说明。对于强智能机器人无法感知刑罚痛苦而言，准确地说，应该是其无法感受人类遭受刑罚时的痛苦。否定论者并不否认人工智能以算法为运行基础，人工智能技术人员根据智能机器人的运行原理将与人类遭受刑罚时痛苦所提炼出的同等剥夺感写入算法，并非天方夜谭。对单位判处罚金时，同样无法让单位体验到人遭受刑罚时的同等痛苦，若按照否定论的上述逻辑，单位作为刑事责任主体的合理性丧失，此时否定论便会面临逻辑上无法自洽引发的矛盾。因而，让非人类主体感受到人的同等痛苦不是核心，核心在于让他们感受到刑罚带来的同等的剥夺感。剥夺感既是刑罚惩罚功能的载体，也是发挥预防功能的保障，应当根据主体的不同进行刑事立法上的相异配置。因而，在以人的认知、感受作为刑事立法尺度评价所有新兴主体时，必然落入封闭的话语体系，使得刑法的开放性遭到扼杀。

2. 质疑二：强智能机器人无法适用“现有”刑罚体系

观点一：权利义务不统一使得强智能机器人难以主张权利。论者以权利义务的统一性作为认定法律主体资格的唯一标准，强智能机器人在面对人工智能创作物被犯罪侵犯时难以主张著作权。另外，人工智能体也难以主张抗辩权、沉默权、回避权等刑事诉讼权利。[16](P73-77)

观点二：强智能机器人不具备受刑能力。否定论主要从两方面切入：一方面，人工智能没有财产权也没有政治权利，附加刑无法施行；另一方面，增设的“删除数据、修改程序、永久销毁”等刑罚执行措施仅是维修手段，无法体现其给人工智能带来的痛苦感。[17](P123)

否定论的上述观点体现出审视问题时方法论的机械性，未能以辩证唯物主义所主张的发展性看待问题，是站在当下看待未来。著名社会学家卢曼曾对高科技发展睿智地谈到，“即使在与高风险技术打交道时，一种与此相匹配的实践也经常导致用先前的经验视角去阐释警告，顾忌风险转换的程度，进而导致忽略这样的警告。”[18](P109)在看待强智能机器人刑事责任这一问题时，同样不能使刑事立法乃至理论交流陷入卢曼上述谈到的“结构漂移”（structural drift）中，应持发展观点。

在否定论看来，权利义务统一论为强智能机器人权利阙如提供了理论支撑。然而该论据也值得商榷。并非法律上的权利就一定对应着同等性质的义务，有时权利就仅是权利、义务就仅是义务，如配合防疫就只是一种法律义务，就业权就仅是一种法律权利。《宪法》有且仅规定了劳动权和受教育权既是权利也是义务，不能人为随意解释权利义务的统一性。再者，刑事诉讼权利能否行使应以主体能否根据社会现实作出决策为标准。在强智能机器人拥有自主意识后，由于如何行使权利的算法已输入运行机制内，强智能机器人在算法规则内行使诸如沉默权、抗辩权、回避权等刑事诉讼权并无运行、沟通上的障碍。另外，否定论还无形之中忽视了刑事辩护制度。即便强智能机器人无法行使上述刑事诉讼权，但只要其有委托辩护人的真实意思表示，不排除辩护人代其行使诉讼权利的可能性。最后，智能机器人拥有财产权只是时间问题。目前世界上已有将智能机器人视为法律主体的适例，享受创作作品的著作权收入。如AI作曲领域的领先公司Aiva Technologies创造了一个AI作曲家Aiva（Artificial Intelligence Virtual Artist，人工智能虚拟艺术家），2017年Aiva通过法国和卢森堡作者权利协会（SACEM）合法注册，成为人工智能领域第一个正式获得世界认可的作曲家，其对自己的作曲作品享有署名权、著作权等权利。[19](P247)在赋予强智能机器人财产权后，其便可以此作为技术升级的资金来源。关于政治权利，其不仅属于法律问题还属于政治话题，赋予强智能机器人在国家高精尖领域相应的政治权利，有助于激发其擅长的数理计算功能，有利于调动强智能机器人的运算积极性。将增设的“删除数据、修改程序、永久销毁”等刑罚执行措施理解为维修手段实乃建立在强智能机器人无法具有自主意识的前提下，若强智能机器人有自主意识，这些所谓的“维修”手段从处罚的严厉性角度不亚于有期徒刑、无期徒刑或死刑。

（二）回应：否定论的不合理性及理由

否定论者对强智能机器人的态度是极其复杂且矛盾的，一方面他们坚信强智能机器人无法产生自主意识，另一方面又认为强智能机器人拥有自主意识后必然会毁灭人类。这其实是否定论者将“人类物竞天择进化中的鼓励斗争、适者生存、好斗的心理状态投射到了人工智能身上。”[20](P164-167)皮格马利翁效应在强智能机器人身上能否实现的关键还是在于自主意识能否人为制造，虽然目前无法实现，但我们对该问题的态度不能表现为新勒德主义，对人工智能引发的新问题应保持倾听和沟通。对强智能机器人刑事责任的否定丝毫不会影响人工智能技术发展的滚滚前轮，只会延误研究的最佳时机。现有研究更多是未雨绸缪的探讨，随着人工智

能技术的发展，理论内容亦会随时保持更新，这一点与刑法教义学注重的解释性是一致的。

否定论的两种反对路径都不知不觉间回归到人类中心主义立场上。若从另外一个角度观察，赋予强智能机器人刑事责任并未偏离人类中心主义立场。以单位犯罪为例，单位本身并非物质可见的实体，将单位视为刑事责任主体有多种考虑，按照否定论的观点，“处罚单位的根据不在于其是否具有自由意志，而仅仅是国家权力对于社会控制和社会治理的需要。将单位作为刑事责任的主体，实质上是国家权力进行社会管理和控制的战略选择。”[21](P159-161)既然自主意识阙如是否定论否定强智能机器人刑事责任时考虑的核心因素，那为何在赋予单位刑事责任主体地位时又将自由意志从考虑的核心因素中剔除出去？此种逻辑冲突是明显的，否定论者难以解释不同主体下判断标准区别化的根据何在。若持该否定论观点，在强智能机器人尚未产生自主意识时，为何不能基于社会控制和社会治理的需要将其视为刑事责任主体？强智能机器人与单位背后都是自然人在实际控制，二者并无不同，为何会得出不同结果？否定论者并未深入说明。否定论者还认为，若承认强智能机器人刑事责任主体地位，则难以解释如下权利义务冲突：强智能机器人生产企业批量“制造人口”违背人类生存繁育的自然法则；无法沿用自然人间的亲属关系；难以获得独立的财产。若照此逻辑，承认某一主体刑事责任主体资格，主体的赋予者就是在制造人口，那单位的注册机构赋予单位主体资格岂非也是在制造人口？是否也需要在母公司与子公司之间演绎人类亲属间的继承关系？上述答案显然是否定的。否定论以单位与强智能机器人的不同来否定后者的刑事责任主体资格，并走不通，反而处处显示出二者并无不同。强智能机器人更强的物质实体性很大程度上说明其比单位更具刑事主体立法的必要性。与此同时，也有一些学者谈到，是不是出现一些新兴科技事物，就需要将其作为刑事责任主体？应当说，并非所有新兴科技事物都有作为刑事责任主体的必要性，一切以能否产生自主意识为核心要素；在新兴科技事物无法产生自主意识的情形下，必然没有作为刑事责任主体的必要性。围绕强智能机器人讨论其刑事责任主体资格，并非在于其属于新兴科技，而在于其未来会出现自主意识。

在本文看来，否定论者以人类中心主义的捍卫者自居，其以“现在”“人”作为判断强智能机器人是否具有刑事责任的标准，无疑使标准滑向形而上学的机械、静止观，亦使刑事理论研究无法为未来的刑事立法提供智识补给。基于强智能机器人的刑事规制必要性，应当从刑事责任层面尝试分析重构路径。

三、技术逻辑下强智能机器人刑事责任的重构路径

学界在讨论强智能机器人刑事责任时一直有一种误解，不乏一些学者认为“讨论”就意味着马上立法，他们主张在智能机器人引发的危险属于“可允许的危险”时由于风险可控而不需要为机器立法，同时在智能机器人产生自主意识并对人类发动战争时由于风险不可控而无立法必要，正是基于正、反方面的论证得出对强智能机器人无超前立法之必要。[22](P159)该观点混淆了刑事责任确认的必要性与确认时间的不同，即便立足于当前弱人工智能时代对强智能机器人的刑事责任进行分析，主要还是集中在刑事责任确认的必要性层面，并非主张即刻立法。鉴于强智能机器人与自然人、单位在自主意识的表现方面并非完全一样，“应当明确强智能机器人在刑事责任与刑罚根据上与人存在差异，需要对刑法理论做出符合强人工智能机器人特征的重新解读。”[23](P118)

（一）刑法内：刑法总论、分论的重构

1. 刑法总论的重构

（1）犯罪论部分

围绕强智能机器人刑事责任的犯罪论重构，现有研究主要集中在刑事责任年龄划分和

权利保留两方面。在刑事责任年龄方面，论者主张以人工智能的智能化程度作为衡量刑事责任年龄的指标依据，由技术人员依据智能评判标准进行技术认定，对于间歇性系统异常或中病毒的人工智能体可按照我国刑法关于精神病人或酗酒的人的状态进行类比。[24](P112)在权利保留方面，论者分别提出政治权利的法律保留、紧急避险权的法律保留（人类生命权优先原则）和自我复制权的法律保留。[25](P249)

毫无疑问，上述观点拓展了强智能机器人刑事责任主体地位肯认的理论空间，但存在主体内涵阙如、刑事责任认定标准不当和权利保留范围过宽的问题。现有研究未能在强智能机器人概念、种类等内涵厘定方面作进一步分析，基于主体内涵阙如问题，笔者建议在现有大一统刑法修正模式下可以在刑法第31条“单位犯罪的处罚”后面增设“第五节机器人犯罪”，同时增设第31条之一、之二对机器人负刑事责任的范围、处罚加以规定。与对单位犯罪的处罚类似，刑法依然保持以处罚自然人犯罪为原则，以处罚单位犯罪和强智能机器人犯罪为例外。在“第五章其它规定”中采取“概念+列举”的方式明确不同种类、级别智能机器人的含义。另外，以智能化程度作为划分智能机器人刑事责任的划分标准具有合理性，与学界认为弱智能机器人无刑事责任能力而仅需讨论强智能机器人刑事责任的理论共识保持了一致。但是，在刑事责任认定标准上单一采用技术认定标准，有失妥当。此时可借鉴精神病人刑事责任能力确认时采取的“医学+司法”的双重鉴定标准，以“技术+司法”鉴定作为认定智能机器人刑事责任能力的标准，以达到技术逻辑与法律逻辑的内在统一，实现双向制约。最后，权利保留方面的观点很大程度上参照了阿西莫夫提出的机器人三大定律，即机器人不得伤害人类个体或者目睹人类个体将遭受危险而袖手不管；机器人必须服从人给予它的命令，当该命令与第一定律冲突时例外；机器人在不违反第一、第二定律的情况下要尽可能保护自己的生存。不难看出，阿西莫夫提出的三大定律主要是为了保护人类生命安全，而赋予智能机器人相应的政治权利，如言论、出版等权利并不必然对人类的生存与发展构成威胁，很大程度上还可以丰富人工智能创造物的数量，对人类不仅无害还十分有利。

（2）刑罚论部分

针对强智能机器人犯罪，有学者建议在刑法中增设删除数据、修改程序、永久销毁等刑罚种类，[26](P134)也有学者将这些措施理解为保安处分而非刑罚。[27](P99)

笔者认为，无论是将上述措施理解为刑罚还是保安处分，核心问题在于未能基于对智能机器人的剥夺感来区分制裁梯度，无法体现罪责刑相适应原则。应当以智能机器人自主意识产生的功能区为据将其区分为硬件、软件与湿件，三者上述顺序所体现的制裁严厉度由轻到重。硬件制裁体现为功能不损坏、自主意识产生的核心算法不丢失，只是对其运行范围等浅层次功能进行剥夺。软件制裁涉及自主意识产生以外的其它核心功能的毁灭，可根据毁灭程度进一步细分为非核心功能部件的部分毁灭与全部毁灭。湿件制裁关涉自主意识产生区中算法的格式化，可进一步细分为降级处理和格式化，降级处理是将强智能机器人降级为弱智能机器人或更低等级机器人，格式化是将强智能机器人从人工智能市场中清除出去。

2. 刑法分论的重构

在罪名增设方面，有学者呼吁增设滥用人工智能罪。[28](P3)这其实是对强智能机器人生产者、使用者的罪名增设，因为在承认强智能机器人具有自主意识的前提下其究竟与自然人犯罪时的罪名有无差异，是一个值得深思的问题。若认为二者之间的犯罪无差异，则无需为强智能机器人单独增设罪名，比照自然人犯罪的罪名处罚即可，只不过需要罪状中明确提示强智能机器人可以构成该罪。若坚持二者犯罪之间存在质的差异，那差异的核心在哪？无形之中会出现强智能机器人自主意识的存在与差异性

立法之间的悖论。笔者主张对强智能机器人犯罪的罪名比照自然人犯罪进行，不过基于处罚成本的考虑，应以对硬件、软件的处罚为原则，以对湿件的处罚为例外，在各种处罚内部又应坚持以轻缓处罚为原则、以严厉处罚为例外。

（二）刑法外：前置法对违法本质的确认

刑法是保障法，“刑事犯罪的危害本质和违法本质，其实取决于前置民商法或前置行政法的规定，而犯罪量的具备，即性质相同的违法行为与犯罪行为的区别界限，则在于作为部门法之后盾与保障而存在于法体系中的刑法的选择与规定。”[29]针对强智能机器人的违法行为与犯罪行为，应确保分别得到前置法与刑法的对应处罚。应当说，强智能机器人违法本质需要从前置法中寻找依据，不能脱离前置法去作罪刑判断。

学界围绕强智能机器人的前置法规制，存在截然不同的两种观点。一种观点以无法在智能机器人内部构建伦理道德规范且采用AI技术管理规范便可达到规制目的，主张无需前置立法；另外一种观点主张区分刑法与前置法的规制范围，却对前置法的规制路径语焉不详。根据刑法与前置法的关系，强智能机器人的前置立法有两种路径：一是单独制定《智能机器人法》，二是将其纳入《治安管理处罚法》体系。两种立法路径各有利弊，单独立法有利于根据智能机器人的特点、种类较为自由地建立违法行为框架，不必受制于现有以“人”为核心的违法行为体系，但会导致立法成本过高，对智能机器人违法行为的理论研究成熟且成体系后方能开展立法准备工作；以《治安管理处罚法》原有体系为基础将智能机器人作为新主体写入，有利于节省立法成本，只需把智能机器人与“人”相比后的异质性描述出来即可，但会导致智能机器人出现新特点后的法律变更速度受制于《治安管理处罚法》原有“人”属体系的制约。笔者主张单独立法模式，坚持实践需求与理论成熟的有效结合原则，恪守单独立法与《治安管理处罚法》中违法行为、处罚体系的有机接轨性，注重智能技术重大变化与法律更新的同步性。在前置法中对智能机器人概念、主体种类、技术更新层次及标准、违法行为表现、处罚种类和等级进行规定，突出对人工智能技术标准的描述和专业技术人员鉴定程序的规范化。

强智能机器人刑事责任体系的重构有赖于刑法内、外共同构建，不仅需要保障刑法内部总则规定与分则罪刑规范内容的体系化，还需要协调好前置法对违法本质确认与刑法对违法行为类型选择之间的关系。

四、结语

在人工智能技术迅猛发展而未来状态难以被准确预测的当下，能否以强智能机器人未到来而放弃思考刑事责任？答案显然是否定的。只有持续关注人工智能技术并对其蕴含的刑事问题进行及时、有效的回应，才能真正地发挥刑法的预防功能。对于强智能机器人未知风险的预测和刑事研究，是人类对人工智能技术刑事风险保持思考和警惕的体现，放弃思考无异于抛弃“救生艇”。

参考文献：

[1]赵竹青. 张钹院士：从技术层面解决人工智能安全问题[EB/OL]. (2020-12-24) [2022-11-15]. http://m.people.cn/n4/2020/1224/c28-14644885.html.

[2]冯丽妃. “技术+约束”突破机器人发展伦理瓶颈：2022世界机器人大会[EB/OL]. (2022-08-25) [2022-11-13]. https://www.cast.org.cn/art/2022/8/25/art_88_195650.html.

[3]中新网. 多国要调查ChatGPT, TA碰了什么红线? [EB/OL]. (2023-04-09) [2023-06-13].http://www.rmzxb.com.cn/c/2023-04-09/3326933.shtmll.

[4]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.

[5]RAY Kurzwell. 奇点临近[M]. 董振华, 李庆诚, 田源, 等, 译. 北京: 机械工业出版社, 2011.

[6]房慧颖. 人工智能犯罪刑事责任归属与认定的教义学展开[J]. 山东社会科学, 2022(4): 142-148.

[7]集智俱乐部. 漫谈人工智能[M]. 北京: 人民邮电出版社, 2015.

[8]乔治·扎卡达基斯. 人类的终极命运：从旧石器时代到人工智能的未来[M]. 陈朝，译. 北京：中信出版社，2017.
[9]默里·沙纳汉. 技术奇点：当机器拥有人性，我们将面对怎样的世界？[M]. 霍斯亮，译. 北京：中信出版社，2016.
[10]冯文杰，李永升. AI刑事责任主体否定论的法理与哲理证成——兼论“人”是什么[J]. 东北大学学报（社会科学版），2020(1)：90-98.
[11]曾粤兴，高正旭. 论人工智能技术的刑法归责路径[J]. 治理研究，2022(3)：113-123.
[12]刘瑞瑞. 人工智能时代背景下的刑事责任主体化资格问题探析[J]. 江汉论坛，2021(11)：105-110.
[13]蒋巍. 人工智能犯罪的主体定位与责任分配问题研究[J]. 广西民族大学学报（哲学社会科学版），2019(5)：199-204.
[14]叶良芳. 人工智能是适格的刑事责任主体吗？[J]. 环球法律评论，2019(4)：67-82.
[15]姚万勤. 对通过新增罪名应对人工智能风险的质疑[J]. 当代法学，2019(3)：3-14.
[16]彭景理. 论人工智能时代智能机器人的刑事责任能力[J]. 大连理工大学学报（社会科学版），2020(2)：71-79.
[17]冀洋. 人工智能时代的刑事责任体系不必重构[J]. 比较法研究，2019(4)：123-137.
[18]骆多，林星成. 人工智能体犯罪主体资格证伪——以刑事责任之实现为视角[J]. 学术交流，2020(1)：104-112.
[19]尼克拉斯·卢曼. 风险社会学[M]. 孙一洲，译. 南宁：广西人民出版社，2020.
[20]王维嘉. 暗知识：机器认知如何颠覆商业和社会[M]. 北京：中信出版社，2019.
[21]赫克托·莱韦斯克. 人工智能的进化：计算机思维离人类心智还有多远？[M]. 王佩，译. 北京：中信出版社，2018.
[22]皮勇. 人工智能刑事法治的基本问题[J]. 比较法研究，2018(5)：149-166.
[23]卢勤忠，何鑫. 强人工智能时代的刑事责任与刑罚理论[J]. 华南师范大学学报（社会科学版），2018(6)：116-124.
[24]马治国，田小楚. 论人工智能体刑法适用之可能性[J]. 华中科技大学学报（社会科学版），2018(2)：108-115.
[25]朱凌珂. 赋予强人工智能法律主体地位的路径与限度[J]. 广东社会科学，2021(5)：240-253.
[26]刘宪权. 人工智能时代的“内忧”“外患”与刑事责任[J]. 东方法学，2018(1)：134-142.
[27]周子实. 强人工智能刑事主体地位的折衷说——阶层论视域下“准主体”的教义学证成[J]. 广西社会科学，2021(8)：99-105.
[28]刘宪权. 人工智能时代的刑事风险与刑法应对[J]. 法商研究，2018(1)：3-11.
[29]田宏杰. 以前置法定性与刑事法定量原则判断行为性质[N]. 检察日报，2019-05-24(3).

【责任编辑　刘绚兮】

Rethinking the Criminal Responsibility of Bottom-Up AI under the Logic of Technology

ZHAO Tianshui

Abstract: At present, the academic community mostly analyzes the criminal responsibility of BOTTOM-UPAI from the perspective of legal logic, and there are few achievements that cut into research from the perspective of technical logic. The emergence of bottom-up AI is a prerequisite for discussing their criminal responsibility, and it is also the main reason for negative criticism of them. The exponential growth and emergence of artificial intelligence technology, military competition, and commercialization have all provided impetus for the emergence of bottom-up AI. In the process of reconstructing the criminal responsibility of bottom-up AI, it is urgent to solve the problems of the lack of subject connotation, improper criminal responsibility determination standards, and excessive scope of rights retention in existing research. The pre legislation for regulating Bottom-Up AI should adopt a separate legislative model.

Keywords: bottom-up AI; technical logic; emergence; anthropocentrism; reconfiguration