重庆大学学报(社会科学版) 2020年第26卷第3期
JOURNAL OF CHONGQING UNIVERSITY(Social Science Edition) Vol. 26 No. 3 2020

Doi:10. 11835/j. issn. 1008-5831. fx. 2019. 03. 001

欢迎按以下格式引用:熊波. 论人工智能刑事风险的体系定位与立法属性[J]. 重庆大学学报(社会科学版),2020(3):142-154. Doi:10. 11835/j. issn. 1008-5831. fx. 2019. 03. 001.

Citation Format: XIONG Bo. On the system orientation and legislative attribute of artificial intelligence criminal risk[J]. Journal of Chongqing University(Social Science Edition), 2020(3):142-154. Doi:10. 11835/j. issn. 1008-5831. fx. 2019. 03. 001.

论人工智能刑事风险的体系定位与立法属性

熊 波

(西南政法大学 法学院,重庆 401120)

摘要:人工智能刑事风险并非属于一种"超个人风险"类型。对人工智能刑事风险认知的主观幻化现象进行逐一诘问,能够得知:超个人风险分为事实层面的现象风险和规范层面的法律风险,智能产品在设计和编制程序范围外,其所实施的严重社会危害性行为仅是一种纯粹事实的现象风险。人工智能产品刑事责任评价的路径阻却在于智能技术本身缺乏生活情感的经验总结、智能产品适用刑罚规范不具备现实意义、深度学习是凭借人类思维模式的基础输出进行的。人工智能刑事风险的立法归责应确立限制从属性,亦即,限制可允许性与超越性的人工智能风险之存在,明确人工智能刑事风险从属于自然人主体。继而,可为人工智能时代刑法立法的科学化探索奠定理论基础。

关键词:人工智能;"超个人风险";智能产品;刑事责任

中图分类号:D924. 3 **文献标志码**:A **文章编号**:1008-5831(2020)03-0142-13

"AI"作为一种概括性科技术语,是由算法系统组成的认知、感应技术。如今尖端从业者倾向于通过智能产品在算法系统中的深度学习,从多层次的集合数据中提取人类的结构特征[1]。由此智能产品自主决策思维模式下的危害行为,开始突显于现实社会的各类境况之中,典型情形主要存在于"自动驾驶汽车、服务机器人"领域。至此,学界开始对人工智能刑事风险的责任承担进行反思,

修回日期:2019-03-01

基金项目:国家社会科学基金项目"刑罚退出机制的价值确立与实践运行研究"(17XFX009);重庆市教育委员会研究生科研创新项目"人工智能刑事风险治理问题研究"(CYS18182);西南政法大学法学院博士生科研创新项目"刑事责任'行政程序前置化'模式研究"(FXY2019011)

作者简介:熊波(1992—),男,江西南昌人,西南政法大学法学院博士研究生,西南政法大学青少年犯罪研究中心助理研究员,主要从事刑法学、科技法学、青少年犯罪学研究,Email:xiongbolawyer@163. com。

进而提出诸多议题。但目前就人工智能刑事风险的性质研究来看,多偏向于一家之言①,其引导社会舆论和学界探究,呈现出刑事责任认定结构的嬗递之特质,从而动摇刑事归责系统的理论根基。理性对待人工智能时代下刑事风险的结构性转变,是否需要动辄更改现有的刑事责任基础,应当立足客观现实的事实基础进行充分的科学论证。盲目地将主观臆造的智能风险引入现实社会之中,无疑大规模诱发、扩散国民无端的恐慌与焦虑。本文拟立足对人工智能刑事风险"超个人化"现象之诘问,从多维视角论述人工智能产品个体化刑事责任评价的路径阻却,并推断人工智能刑事风险具有人类危害行为的依附性(拟制性)和决定性,以便立足既有的刑事责任理论,理性对待智能科技风险。

一、现象诘问:人工智能刑事风险是否呈现"超个人化"

随着智能时代的新技术和新工具的不断涌现,数字革命历经从"优化组合"到"模块分析"的演进,最后蜕变为"自主决策"的阶段化学习。在智能科技渐趋成熟之际,人工智能产品创设了一个模仿或复制个体"人"的思维过程[2]。在自主学习过程中,人工智能产品可能会独立脱离技术初始研发者、使用者和管理者输出的基础思维,继而操作一系列严重危害社会的行为。对于智能产品的现实行为危害,是否有必要单独依据运行基础系统的算法模式,将其理解为一种"超个人风险"类型?"就一元论的先验个体图式而言,在这一理论支配下的超个体法益样态,虽然摆脱了经验式个人的束缚,但又走向了另一极端,即否定了社会及身处其中的个人之真实性"[3]。因而,承认超个人法益的学者亦然妥协地表明:"超个人"实质内涵的理解虽然挣脱了经验式个人的桎梏,但与此同时,其也忽视了人之所以为"社会人"表达的真实性。机器学习具备一定基础的思维训练,但在视觉、听觉、语言方面的系统处理过程中,智能产品可能在设计者的编程和程序外产生独立的行为自由。

由此导致的现实问题之一:是否意味着通过后续的深度学习,"智能机器人完全可能具有与人一样的辨认能力和控制能力,也完全可能通过深度学习,在设计和编制的程序范围外实施严重危害社会的行为"[4],继而认为其是一种人工智能产品的超个人风险类型,据以赋予它独立承担刑事责任的能力,将超个人风险评价为刑事法约束的风险类型?刑事责任能力作为非难可能的一种人格状态,其要求对于超个人风险的罪责评价,应当是构建于行为要素的理解力和刑罚的承受力基础之上的。因此,如何评价人类基础思维灌输后的智能产品深度学习的超个人风险,则成为化解编程和系统外的独立危害行为的关键性问题。犯罪事实作为刑罚苛责的基本前提,应当明确法律事实和客观事实的评价差异[5],将超个人风险划分为事实层面的现象风险和规范层面的法律风险。智能产品在设计和编制程序范围外实施的严重社会危害性行为,则是一种纯粹事实的现象风险。主要原因在于:智能产品缺乏对危害行为的理解力和刑罚的承受力,在危害行为发生之际,智能产品并无能力提前预知行为风险的存在,进而控制和支配危害行为的进程。

此外,刑罚的承受力无法在人工智能产品本身得以疏通,销毁机器、删除数据、更改系统等针对性刑罚设置无异于多此一举。刑罚苛责对于机器人而言,仅是一种外在数据信号的映射,一种同外界环境变化的现象感应。鉴于此,超个人风险规范层面的法律风险类型,无法在智能产品本身中

①如下文将要论述的智能产品"独立刑事责任论""权利主体类型正当性"这两种典型观点。

探寻。

然而,现实问题之二:机器人系统操作的一种纯粹辨认能力的根源何在?算法作为人工智能系统运行的基本路径,智能产品一旦牵涉“意识”,强人工智能的定义和评估标准就会变得异常复杂。“拥有意识的机器会不会甘愿为人类服务?机器会不会因为某种共同诉求而联合起来站在人类的对立面?一旦拥有意识的强人工智能得以实现,这些问题将直接成为人类面临的现实挑战”[6]。而法律层面的挑战莫过于机器人本身的行为危害。机器人“意识”来源于外界环境变化的接收,算法技术自主调配数据的决策分析过程,则是操作系列行为模式的塑形。算法技术在操控智能产品危害行为过程中发挥着举足轻重的作用,“其包括的设计、目的、成功标准、数据使用等都是设计者、开发者的主观选择,它们可能将自己的偏见嵌入算法系统。此外,数据的有效性、准确性,也会影响整个算法决策和预测的准确性”[7]244。换言之,机器人系统操作的一种纯粹辨认能力的根源在于人类行为本身。对于智能产品危害行为的自主操作,仅是超个人风险事实层面的现象风险类型。

在智能化时代,客观行为造就的风险类型样态多元且构造复杂。因而,现实问题之三:如何区分理解人类智能产品的滥用行为?防止将无责任非难基础的智能风险划归为行为人。毋庸置疑,人工智能刑事风险的不明确性或者不确定性,致使“预防原则”仍应当是风险化解的主要思路。但是,风险刑法的罪名设置必然要经受两个基本问题的考验:一是人工智能刑事风险在传统刑事法体系中如何定位?二是刑法化解人工智能刑事风险的依据是否合理?随着自动化取代了更多的劳动形式,创造性的表达似乎成为现实必然。机器无论如何发展,都是人类技术规划的意识反映[8]。因而,适度承认科技社会发展中现实存在且能够承受的“允许性危险”,可在助益生产力进步的同时,保障国民对刑法体系的基本认同感,实现刑事立法的科学化探索。

二、体系定位:人工智能机器刑事责任评价的路径阻却

人工智能刑事风险的精准定性问题,决定着刑事责任评价主体的范围。出于实践经验的价值遵循,智能产品能否拥有独立的刑事责任能力,继而被认定为刑事责任主体新类型,主要在于对法律事实的辨认能力、控制能力。现阶段部分学者大力支持的智能产品“独立刑事责任论”以及“权利主体类型正当性”的观点②,完全脱离法律的首要泉源,违背人工智能技术刑事规制的必要目的与手段的正当性等要求。

(一)智能技术本身缺乏生活情感的经验总结

智能技术的日臻完善与人类情感的价值流变,使得智能机器程序产生“冷精神”或称“系统精神”意识。人类理性与机器意识,主要的区别在于危害行为的方式选择。前者是出于朴素生活情感的经验表达;后者是自主决策意志的模块反映。刑事责任作为一种行为否定性评价的法律后果承担,应当关注“以存在论和价值论作为人性分析框架”的适用范围[9]。强行扭曲刑罚的社会基础,将人工智能产品视为一种非生命体的自由意志表达者,进而认为机器在刑事责任承担方面,其唯一差异仅在于生理构造[10],完全是一种“超个人化”风险的观点。一方面,个体能够清晰地识别风险,判

②观点详情可参见: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》(《法学》2018年第1期,第40-47页);Radutniy O E. Criminal Liability of the Artificial Intelligence(Problems of Legality,2017(1):132-141)。

断风险对自身利益的影响以及攫取自身利益最大化的保全方式。另一方面,为了自身利益的集中化采取各种手段逃避风险,推卸责任,从而无法达成有效的集体行动。将智能产品行为视为超个体风险的法益侵害,虽是“保障内在自由(绝对命令) 外化的不可或缺的体制”[3],但这一概念的塑造,完全将刑法危害行为的控制能力视为一种脱离经验和价值判断的程序精神,因而,其仅具有形式意义,而不具有实质的意志指向性。

1. 人工智能的自主决策仅是一种纯粹事实行为,并不具备责任非难可能性

刑事责任的可行性基础在于责任主义,即只有能够非难时才可科处刑罚。因而,非难可能性作为刑事责任评价的基本前提,要求刑罚否定结果的正当性和合比例性。正当性在于刑事责任的承担是智能产品对事实行为的经验选择,是在全面熟知行为举动的发生之后,对宏观、微观双层环境的整体变化之作用力,其具备的情感经历是对刑罚“质之变化”的决定要素。而合比例性反映在生活阅历之上,对责任非难可能性的认识程度,最终影响刑罚苛责的“量之变化”。对于人工智能产品执行过程的高度复杂性任务,机器人不断增长的执行能力和自主权的能力,是建立于数据处理系统基础之上的。

质言之,对于瞬时间的行为危害的产生,机器人也仅是意识到该行为是在接受数据信息后的系统操作,而对于系统操作的潜意识理解是缺乏的。由于智能产品缺乏对情感生活要素的理解和体验,因而,危害行为的责任非难可能性因环境人格之缺陷而无从谈及。知识来源于感觉经验,在认识上人的心智受感觉影响,感觉因物质事物而起[11]。法律立足机器责任视角,对人工智能产品的危害行为评价,仅能将其评价为一种自主决策系统反映的纯粹事实行为,而无法强制纳为刑事责任的非难可能性。

2. 人工智能的危害行为仅是系统语言的模式选择,并未受到主观心态支配

智能机器人能否单独承担刑事责任,除实施客观的危害行为之外,还包括受到主观心态的支配[12]。毋庸置疑,人工智能的系统操作完全依赖于算法系统的模式选择,换言之,技术设计和编制程序内外行为的发生,都依赖于算法系统运作的好坏。因而,在此方面,智能危害行为的意志支配完全取决于算法系统的机械性、自主性运作。语言系统作为智能产品的程序精神表达,语言代码的编排很大程度上受制于客观环境和地域文化的影响。语言作为生活情感的表达方式,与生存环境的变化密切相关③。智能产品无法用系统语言表示“我为何要实施犯罪行为”,理想超前,但基础设施尚在襁褓中的人工智能将面临一大问题:算法逻辑自身的问题[13]。数学方法的发展还不够,加之硬件计算能力的不足,算法语言的错误无法避免。比如,机器翻译的语法错误就是典型问题。科学家夜以继日地总结人类语法规则,设计计算机语言模型,机器却始终无法把翻译准确率提升到令人满意的程度。因为语法本身就是一个用语习惯和地域环境自然而然形成的[14],必然是系统语言在主观罪过因素上无法逾越的罪责鸿沟。

3. 人工智能危害行为的自由意志仅是一种程序精神,并不涵盖实质的辨认能力

回归刑事责任的道义基础,人工智能产品需承担独立的刑事责任,必须论证其具备齐全辨认能

③同一个国家的不同区域,语言也会因环境气候和人文习俗的差异而存在特质性,致使人们对不同地区的话语形式难以及时理解和吸收,如地区方言的形成。

力和控制能力。辨认能力作为控制能力的前提条件,作为超越技术控制和意志自由的“强人工智能”,是否能够将机脑中代码程序的信息集合和数据决策,作为刑事责任能力的辨认要素,仍有待商榷。有学者认为,“强人工智能产品的认识水平,在很大程度上会被局限于特定领域,而缺乏对社会生活的整体认识”[15]。但是,决意实施善恶行为之前的意识判断本身就是一种生活经验的高度融合,强人工智能产品也依托于现实环境生存,但这一依托仅限于外在形体,内在机脑算法系统的深度学习行为,本质上仍是一种程序精神使然,与刑事责任的辨认能力之本质属性具有天壤之别。

(二)智能产品适用刑罚规范不具备现实意义

即使部分学者承认智能产品能够依据算法程序,脱离人类思维模式的基本控制,单独在自主意志支配下,实施人类设计和编制的程序范围外的犯罪行为,但这亦需要经受刑罚归责的必要性和可行性之考究[4]。

1. 智能产品缺乏刑罚的识别能力和承受能力

由于人类科学需求的多元性和适用环境的复杂性,目前关于强人工智能意识、思维发展的讨论,是建立在人类科技总是以加速度形式跃进的基础之上[16]。科技总能超越人类生理结构的发展速度,但是科技产品对事物的认识能力和对痛苦的承受能力,是否能确保与人类同等的精确,是需要在客观环境的历练中感同身受的。刑罚学理论中关于刑罚功能的分析,如惩罚的功能、改造的功能。究其源头,都是基于社会学的功能理论。因为,任何行动系统都需要在适应外界环境的前提条件下,确定一定的行为目标,并在该目标指引下开展行动,将行为系统的各个部分基于功能最大化的要求,进行优化组合[17]。鉴于此,智能产品对刑罚的识别能力,应当体现在外界环境的适用功能建立的基础上。如果缺乏外界生活环境的阶段化长久感知,如何辨别好坏,避免触及刑罚,便成为机器人负刑事责任的一大阻碍。

在社会活动的实践操作方面,具备资格能力的社会成员都是社会关系的物化。承受能力作为一种对刑罚适用最高强度的感知,要求建立在社会活动的现实操作基础上去感受刑罚的强度。因智能产品的操作熟练所造成的严重危害行为,仅是基于数据自主决策分析的瞬息性、高效性,而并非出于对事物强度认识的整体把握。智能产品对量刑基准规范与价值的实质理解,无法在人工智能先占性发展战略的要求下获得。因为,“对一个人类的观察者而言,对机器人的物理损坏或者损害可能看起来像是肉体惩罚甚或死刑;但是,它对机器人并不具有类似的影响,至少它不具有(身体健康地)活下去的意愿”[18]340。所以销毁机器、删除数据以及更改配置等刑种配置及其程度差异,对强人工智能产品的自主意识而言,简直是天方夜谭。

2. 刑罚适用于智能产品无法获取规范公允力

发挥刑罚适用的正当性基础过程,是一个刑法司法文明的精神体现。“无论是刑罚的一般预防还是特殊预防,都要求必须在不破坏‘社会情感’的程度上才能实行”[19]。社会情感渗透于刑事司法程序,是对刑罚文明的公正性、妥当性的一种朴素性色彩的要求。在社会基本认同的情感中,我们即使承认未来科技能够运用模拟仿生手段,使机器神经网络的运作模式与人脑极为相似,甚至超越人脑的思考空间,但也并不代表公众对“智能产品如同人类主体一样,具备刑罚的独立承受能

力”④等类似观点的普遍认同。

诚如,在人工智能技术产业发展的前景预测中,智能科技学者谈及“神经网络类似人类大脑,由一个个神经元组成,每个神经元和多个其他神经元连接,形成网状”[20]。但是,“类似、仿生”人脑始终无法做到与人类思考模式、生理神经相等同。“刑罚在一定条件下被用作凝聚群体,强化共同价值、共同信念的仪式”[21],对刑罚适用的功能最初认识(如报应刑理念)是出于强调对道德伦理秩序维护的自然法所体现的一种朴素情感。对智能机器施加刑罚,承认机器的刑事责任主体地位,无法在特殊群体中获取认同感,亦无法在社会人群中调动或增强守法意识。

3. 刑罚适用归根结底是对自然人的现实惩罚

对智能产品适用新类型刑罚仅是一种抽象概念的意向性表达,即意在借助刑法立法的新思维,说明罪责自负原则中的“超个人风险”的现象类型,应当归为拥有深度学习、自主决策的强人工智能。但从本质而言,智能产品的刑事责任适用归根结底是对自然人的现实惩罚。虽然发现智能产品脱离人类的支配控制,进而实施了危害行为,理所应当将责任归于机器,表面上看,合乎情理,但实质上在刑罚法理基础上无法疏通。由于一个机器人没有财产(至少,它不可能知道它拥有财产),任何判处罚金将不得不由其法定的所有人或者为可能的责任而创设的特定基金来支付[18]347。纵使智能机器人具备独立意识,但其对人类仍然存在强烈的人身依附性。对其新设“删除数据、修改程序、永久销毁”等刑罚种类[22],在实际效果上更类似于对人类“犯罪工具”或财产的没收、销毁,而这本质上是对人类自身的惩罚。

(三)深度学习凭借人类思维模式的基础输出

国家研发、使用战略层面,将人工智能定位为“研究、开发用于模拟、延伸和扩展人的智能的理论、方法、技术及应用系统的一门新的技术科学”[7]23。未来人工智能产品的尖端研发,至始至终都无法脱离人类思维模式的基础输出进行。机脑的塑成本身是一个人脑的高度仿生建模过程,强调“理解、建模和模拟人类大脑的关键是对大脑新皮质实施逆向工程,而大脑新皮质是我们进行循环分层思考的部位”[23]。因此,深度学习的物质载体包括人脑的意志性规划。

1. 思维本源:研发者的编程设计

首先,在技术发展层面,我们不知道强于人类的智慧形式将是怎样的一种存在。目前,强人工智能具体发展的形态、规模以及思考模式都无法实现精准预测。现在谈论强人工智能和人类群体的关系,仅能将智能产品视为一种工具与手段,强人工智能产品必须依照人类的基础思维模式去运作。在智能技术的风险预测和防控范围内,出现的现实危害结果应当归为人类的一种责任疏忽。然而,就现在的科技水平预测强人工智能产品完全独立的操作模式,根本不存在可以清晰界定的讨论对象。其次,在法律规制层面,人类出于社会集中化利益的考量,对智能时代的物质把握,节制在可掌控的程度之上。尤其是在潜在的科技研发道路上,人类群体的科技利用应当考虑在法律约束范围内。即使是在高度自动化后工业时代,人类也需要将操作指令输入信息系统,并辅之基本的监督和维护。因而,在技术层面尚未确定强人工智能的发展规划和基本定义等一系列前提下,强人工智能后续深度学习的运作模式或行为操控,仍取决于研发者的编程设计水平。

---

④观点可参见:刘宪权《人工智能时代刑事责任与刑罚体系的重构》(《政治与法律》2018 年第 3 期,第 89-99 页)。

2.后续学习:使用者的思维输出

支持智能产品“权利主体论”的一派学者认为,机器人自从被创制后,就具有一种独立学习的行为能力。因而,在某种程度上其会导致权利主体、权利形态和权利适用方式相继发生变更[24]。但是,其忽视了人类思维模式的基础输出,并非仅在智能产品技术研发和生产阶段才能实现。智能产品的后期深度学习,仍然受制于使用者、管理者的思维掌控。虽然机器挑战、培养所谓的“超级任务”意识,但是出于人类行为目的的自洽,挑战和培育仍是按照远程操作的使用者给出的既有指令,有步骤依次完成的。而对于完成过程中出现的意外事件,仅是学者过度解读深度学习功能,认为使用者基础思维模式外的危害行为,已然脱离了除机器本身以外的任何思想支配。但殊不知,智能产品深度学习的自主决策行为或意识,完全是按照使用者、管理者的行为目的实施的,受行为人基础思维模式控制。

三、性质证立:人工智能刑事风险归责的依附性与限制性

既然人工智能刑事风险不具备将刑事责任苛责智能机器独立个体的机能,那么着眼于未来的全智能化时代,必须找寻制度构建的科学依据并确立人工智能刑事风险的立法性质。

(一)附属性:附属于自然人

当前,智能科技的普及化运用必然会成为新时代的显著标志,科技风险作为典型的内生性风险,“为应对现代社会的内生性风险,出于控制风险、一般预防的目的,刑事立法逐渐出现了法益保护早期化现象”[25]。在前置刑法预防的现象下,笔者认为,类型化审慎评析人工智能的刑事风险,便可察觉其存在着“行为、主观”这二类科技风险附属于人类。

1.行为附属

出于对犯罪行为特质性的多元视角考察,刑法上的行为概念通常包含着因果行为论、目的行为论、社会行为论和人格行为论四种学说。因果行为论表明犯罪行为是一种有意行为,是在意识支配下表现于外的肢体动静[26]62;目的行为论在因果关系论的基础上发展而来,其认为应当将行为理解为“有计划地使因果发生的要素,在行为的‘目的性’中发现了一个十分重要的连接点”的客观表现[27];而社会行为论揭示行为的外在举动,认为“刑法是一种社会统制(同统治)手段,故具备社会意义的人的身体动静才是刑法上的行为”[26]63。人格行为论者更为直截了当地指明,行为是“行为者人格主体的现实化”,身体动静之所以加以刑事归责,是考究人格与环境的相互作用下人格责任的最终结果⑤。

由此可知,社会行为论、因果行为论、目的行为论,以及人格行为论分别将“意识支配、目的规划、人格环境”作为行为的原因力,均是自然人主观意志状态的现实化结果。进而应当将智能产品的媒介主体认定为一种行为工具,或是间接正犯的行为支配对象。如此,自然而然刑事归责理论才能得以顺畅。智能机器的超脱编程和设计程序范围内的自主违法行为,可以被解释为一种不法而无责的无刑事责任能力的行为。

---

⑤详细内容可参见:[德]汉斯·韦尔策尔《目的行为论导论——刑法理论的新图景》(陈璇译,中国人民大学出版社,2015年版第1页);[日]大塚仁《刑法概说(总论)》(冯军译,中国人民大学出版社,2009年版第106页)。

2. 主观附属

之所以将责任非难可能性与刑事责任的承担关联起来，旨在表明行为主体的不法谴责，是建立在行为主体能够认识到其行为的内容、社会意义与危害结果的基础之上。由于生活情感和社会体验的匮乏，再加之行为进程的推进受自然人支配，智能产品的违法性认识亦应当附属于自然人。正如前文论述，智能产品的技术研发和深度学习阶段，是凭借人类思维模式的基础输出进行的。然而，在智能产品实施危害行为之际，客观行为主体是如何认识到或者可能认识到行为的法规范违反呢？厘清智能产品刑事主观风险的主体归属性，应当明确主观答责的故意和过失要素。

其一，主观故意的人类附属性表现在隐密于智能产品背后的自然人可以全盘掌握并促使明知行为样态和结果类型的发生。智能产品无法将希望和放任的主观意志表达出来，因为在本源上便不具备积极追求或者放任结果发生，结果也并非实施行为直接、间接追求的结局。而自然人可以通过借助智能产品的工具行为或者直接行为，达到主观故意的危害结果实现。但是，在正常范围内的研发和使用，出于智能产业发展的战略需要，即使明知深度学习期间可能发生的突发系统错乱和数据调取失范的行为，亦应当被视为一种可允许性科技风险。

其二，主观过失的人类附属性主要表现在，作为设计者和使用者，本身基于科技使用的目的而创设一系列智能产品，应当提前预知到会由于缺乏必要的全面检测和技术规范，而导致后续的脱轨行为。对于预见义务的履行，智能产品无法分辨出预见事物的本质，或者在研发阶段缺乏表明系统制作过程中的数据错乱和漏洞存在的客观可能性。主要原因在于：一是基于信息环境的流动性和共享性，智能机器对于收集信息本身的性质无法及时甄别和提前预知；二是囿于生存环境的复杂性，智能机器无法依据深度学习获取危险环境的观察判断力。

在实证科学看来，环境风险性质的预判力是一种实证性规范[28]。过失犯设置的目的在于苛求行为人严格按照自身的知能水平和规范能力履行结果回避义务。因为每个人的认知能力存在着显著的差异，如果忽略人类思维模式基础输出以及后续的监督过失行为，为千遍一律的智能机器单独设置一套新的刑罚体系，以应对智能刑事风险，即使在客观上惩处了机器人，但实际上，当前智能产品并不拥有自我反省的能力。

（二）决定性：受自然人控制

不同于上述人工智能的行为风险与主观风险附属于自然人，其旨在表明虽然客观形式上，强人工智能的编制和设计程序范围外的危害行为和自主决策思维，在深度学习过程中，是智能产品单独实施和拥有的，但实际上其可拟制为人类行为的一部分[29]。而本部分论述人类对智能产品的控制，主要表现在部分强人工智能危害行为的发生，完全符合人工智能刑事风险的人类自我罪责范围；技术的研发和危害结果的衍生过程，完全受自然人的意志决定。其实施的行为本就应当视为人类实施的客观行为，而并非是一种拟制的自然人行为。

1. 技术的研发进程受自然人控制

立足国家政策层面，诸如人工智能此类新兴科技趋势的推动，完全得益于政府科技创新的前瞻性引领与控制。从《新一代人工智能发展规划》到《促进新一代人工智能产业发展三年行动计划》，人工智能的先占性、先机性战略的奠基，为科学前瞻预防和约束风险扩张铺垫了良好的发展态势。

政府在风险规制决策的环节,始终无法挣脱“平衡防控风险以保障安全和允许创新而发展科技这两种对立要求”之漩涡[30],因而零散化、碎片化的危害行为之刑事可罚性,无从苛责技术人员和研发人员。

但是,从人工智能全篇布局的整体规划中,我们不可全盘否决自然人在持续支配着技术研发成果和智能产品操作系统的进程。因而,在前述故意和过失主观罪过层面,将自然人在研发过程中因故意或者疏忽心态导致系统漏洞的行为,评价为行为人的自身实施的危害行为,完全是出于智能机器作为行为载体的一种表现形态。因此,研发制造的法益侵害本身的创设,也可以被评价为一种人类犯罪的表现形式⑥。美国科幻作家阿西莫夫曾提出“机器人三原则”,即“机器人不得危害人类、必须服从人类的命令、在不违反第一条和第二条的情况下必须保护自己”[31],亦可侧面表明技术的研发行为实质上是一种命令的创造行为。

2. 危害结果的衍生过程受到支配

在研发阶段终结后,面临的则是弱人工智能指令输出过程和强人工智能的自我决策过程。通常而言,弱人工智能使用者的指令输入行为,作为一种典型的支配型犯罪,其作用力的全程支配为危害结果的罪责实现提供了刑事可罚性的路径,亦在责任非难可能性和刑事违法性的双向结合层面,为刑罚适用提供现有体系基础[22]。但是,强人工智能深度学习情况下的行为超越,仅是如前文所述,被拟制为一种人类附属性行为吗?其实不然,深度学习阶段的危害结果的衍生过程,同样能直接评价为一种自然人犯罪行为。譬如,智能产品的使用者将受害者的日常行程的路线等详细信息,植入强人工智能机脑之中,意图通过大数据信息的智能整合来实现杀人的目的。

其实,强人工智能的不利情况排除,促使行为人杀人计划顺利进行,便是危害结果衍生过程的行为人支配结果。因而,前期的强人工智能对犯罪环境和危害结果的发展进程的数据分析,应当视为故意杀人罪的一种预备行为,而并非前述的借用强人工智能之手来侵犯法益。众所周知,任何立法都不是制定者的主观臆想的表露,尤其是在智能技术时代,刑事立法更应由现有的社会条件所决定,使之与实践发展相适应。从人类拟制行为到直接行为的实施,都足以表明,智能产品的特定危害行为的刑事可罚性机理,在于人类对结果衍生过程的支配力。而正是支配过程的助益,使得罪责自负的主体评价范围仅限于自然人,而并未涵盖智能产品。

3. 人类不可能允许技术超越伦理

刑法的目的和任务在于保护法益,法益作为人类所持有的各种合法利益,人类对于智能产品的所有权,可以被视为刑法法益的一种类型。智能产品尽管被作为社会机器人用来与人互动,但就其根本而言,它们也不外乎是各种不同技术原件——诸如机箱、磁板和传感器的组合。在现有的立法体系以及刑罚归责路径阻却下,将机器人作为刑事责任主体,有违人类对基本伦理的现有认知。立足对合成智能的价值审度,我们应当呼唤更加透明、可理解和可追责的智能系统,强调算法决策与算法权力的公正性,努力调和或化解智能化技术的合理性以及人的主体性存在价值的冲突等诉求。

基于社会秩序的维护和基本生活的满足来源于物质发展,机器人作为基本生活和物质发展文

⑥具体参见:山口厚《刑法总论》(付立庆译,中国人民大学出版社,2018年版第45页);松宫孝明《刑法总论讲义》(钱叶六译,中国人民大学出版社,2013年版第45页)。

明的一部分,难以同自然人主体并存,共同体会基本道德和伦理对自我行为的约束力。况且,人类将刑法作为制度、权利保障的最后手段,因此,“不切实际的、未实现的未来概念,比如机器人杀手?我们根本无法确切知道杀手将要怎么做。但是大多数情况下我们都在试图预测不存在的超系统,在可预见的未来预测可能不允许存在的东西,但这不是法律通常的发展方式”[32]。机器人单个行为主体的认定,首先会在基本伦理范畴受到强大阻碍,何况作为后盾法的刑法保护。人类行使权利能够享受正义的喜悦,而脱离正义观念或者将正义价值附属于自然人,去认识机器人权利主体地位,完全背离伦理的认知和常识。

(三)超越性:可允许性风险

面对智能产品危害行为属性的渐变和形态的多样,持有智能产品刑事责任主体论的学者认为,将智能风险全部归为自然人,有违刑法的人性化处断和理性刑事司法之境遇。为此,在此基础上,将机器评价为刑事责任主体,可以部分涤除智能科技的固有缺陷。但是,不容置疑的是,并非所有的科技风险都是刑法的评价对象。目前存在的智能科技风险扩张化趋势有其特定的背景支撑,一方面在国家层面上,源于科技专家和政府部门对智能产品未来发展方向的模糊化和不确信性,激增国民整体的不安全感,其诱发刑法前置化预防理念再次显现,以便为超前性刑事立法提供制度保障;另一方面在国民层面上,基于民众对科技产业的满怀期待和过度认知,徒增科技风险的潜在心理威胁。人工智能风险的不明确性,致使犯罪被害的感受愈发强烈,刑法偏好在某种程度上影响刑事立法方向[33]。因此,立足科技固有的尝试性发展路线,化解科技专家对智能风险评估与公众智能风险认知之间存在巨大差异的关键性问题,在于倡导可允许性的智能风险规则[34]。本质上,在借助本文主张的智能产品行为的附属性理念和人类绝对控制性理念之后,机器编制和设计程序外的自主行为,仍超脱人类思维意志范围,那么该种风险便是法律上可允许性风险。其法理根基就在于,部分人工智能的客观风险是人类社会无法预知、无法规避以及无法规制的科技附属品。

1. 无法预知

行为人能够在某种程度上主观预见到人工智能刑事风险的存在,则表明其应当承担防止预见风险实现的义务。如果违反危险排除义务性的强制规定,滥用或放任可预知的人工智能风险,致使其肆意侵蚀国民法益,行为人则应当为其可预见的主观罪责风险负相应的刑事责任。然而,判定无法预知的智能风险的辐射范围,不应当按照国民的基本认知水平予以界定,而应当依据相当性原理进行判断。具言之,以禁止性危险作为客观归责的前提条件,对于预见的因果关系判断,应当以社会经验法则的具体标准进行界定,并结合行为构成要件模式加以区分判断[35]。

“无法预知”强调虽然智能产品危害行为附属于或者直接受行为人支配,但是在举动前行为人无法预知风险结果的发生,抑或是在行为时无法预测结果的产生与发展(包括结果的有无、具体形态及其发展方向),则不可轻易将客观科技风险强加于自然人。此外,行为、结果的发展过程会因相应犯罪构成要件内容的特殊性,而影响相当性原则的判断。譬如,身份犯的具体职责性要求、结果犯和行为犯的既遂标准。因而,笔者认为,采取相当性理念必须摒弃固有的一般人的思维模式。对于研发者、使用者和管理者附属或支配的危害结果的预知能力,相当性理念应当结合构成要件的特殊性,采取不同的标准进行解读。详言之,研发者能够按照现有的技术性操作去设计智能产品,使

用者能够按照出厂的严格程序和方法使用和保养智能产品,管理者能够及时承担排除缺陷的智能产品义务。否则,在上述范围外的危害行为或者结果产生,都应当归为无法预知的可允许性智能风险。

2. 无法规避

人工智能技术日益复杂的信息表达,更新数据化时代的信息传播、收集、合成等手段方式。智能信息表达实现了人类对计算机化服务和产品的空前掌握,以迥异的交流方式体现了相当范围的交流支配权。因而,纵使技术条件日益更新和完善,我们仍拥有前瞻性的智能风险预测能力,否则放任无法预知的风险进入现实社会,那么在刑法体系整体崩塌前,正常的社会秩序必然首先受到不可控制的大规模风险侵蚀。清晰且明确塑造技术信息的不法规则,应当排除已经预见但无法及时规避的人工智能风险。而且这种智能风险应当仅限于特定阶段,基于数字危险不当进入现实社会的防控需求。

特定阶段的认定既然是出于虚拟风险的技术规制需要,那么在刑事法律层面,我们应当允许此阶段所发生的现实危害。诸如,自动驾驶汽车的技术测试环节,实质上就是检验预知风险存在真实性的一个阶段。因而,刑法上的可允许性风险应当可以明确在智能产品测试环境下,研发者和使用者拥有的仅是部分风险预知能力。针对测试环节发生的已经预见的人工智能风险,在依据现有的技术能够规避的情况下,我们则需要将其视为一种智能科技的禁止性风险。譬如,测试者或使用者明知研发阶段或者测试前阶段,智能产品环境感应已经发生故障,不进行技术排除仍放任测试而导致危险发生。如果在规制风险的制度构建层面,不加限制可允许性智能风险的认定范围,尤其是在已经预知但人类无法规避的测试阶段,那么自然人可以改造和创制的技术风险将不受约束。

3. 无法规制

既然人工智能风险在技术层面上存在着无法预知、无法规避的纯粹事实风险类型,那么在法律层面上就无法将上述不具备期待可能性或者是业务职责行为纳入客观归责体系之中,本质上此阶段的强人工智能所实施的严重社会危害性行为,则是超个人风险类型中的一种纯粹事实的现象风险,其并不属于刑法规制的智能产品风险范畴。通常而言,行为人虽然制造了现实所不允许存在的危险,但出于客观情境和刑事法的强制性之考量,该危险是为刑法所允许的危险时,则排除结果之归属,行为人不需要为该结果负责[36]。

基于对“算法具有难以预料、无意识的属性,而并非编程人员有意识的选择,其增加了识别问题根源或者解释问题的难度”的考虑[7]243,秉承人工智能独立刑事责任论的学者亦持“限制说”的观点[14]。因此,撇开智能科技建构社会的技术层面不谈,在全部人类建构生活模式所依据的生活资讯项目当中,在涉及刑法所关心之利益思考上,智能风险内容固然也仅指涉及个别犯罪事件。

四、结语

在人工智能犯罪主体单独化理论研讨路径受到阻却之后,可知人工智能刑事风险是“超个人风险”类型中的一种纯粹事实现象,其并不属于智能产品本身所应承担的独立刑事责任范畴。人工智能产品行为,要么可附属于人类,要么直接受自然人控制,因而,刑事违法性和责任非难可能性的评

价基础,仍在于自然人主体。除此之外,人工智能刑事风险在排除客观归责中的可允许性风险之后,即将面临运用现有的刑事法制度予以规范和约束的命运。对可罚性的人工智能风险进行刑事责任评价,是一个运用法律文本的社会控制行为。“通过法律的社会控制实际上是通过语言的社会控制,无论是立法上的空间展开还是司法上的时间运动,法律文本和法律运行决不可能脱离语言”[37]。因而,借助刑法教义学对人工智能刑事风险进行介评,关键在于现有刑法规范语言的妥当适用与技术性规范缺失的制度完善[38]。

参考文献:

[1]CALO R. Artificial intelligence policy:A primer and roadmap[J]. U. C. Davis Law Review,2017,51:399-436.

[2]ABELLERA R,BULUSU L. Oracle business intelligence with machine learning[M]. Berkeley,CA: Apress,2018.

[3]贾健.人类图像与刑法中的超个人法益:以自由主义和社群主义为视角[J].法制与社会发展,2015,21(6):127-140.

[4]刘宪权.人工智能时代的刑事风险与刑法应对[J].法商研究, 2018,35(1):3-11.

[5]陈伟,熊波.罪数形态中行为定量分析的理论重构:兼对“行为”立法模态化用语之辩正[J].西南政法大学学报, 2017, 19(2): 54-66.

[6]李开复,王咏刚.人工智能[M].北京:文化发展出版社,2017:118.

[7]腾讯研究院,中国信通院互联网法律研究中心.人工智能:国家人工智能战略行动抓手[M].北京:中国人民大学出版社,2017.

[8]NAUČIUS M. Should fully autonomous artificial intelligence systems be granted legal capacity? [J]. Law Review,2018,17(1):113-132.

[9]陈兴良.刑法的人性基础[M].2版.北京:中国人民大学出版社,2017:336.

[10]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):40-47.

[11]白建军.刑法规律与量刑实践:刑法现象的大样本考察[M].北京:北京大学出版社,2011: 55.

[12]BROWN D K. Criminal law reform and the persistence of strict liability[J]. Duke Law Journal,2012,62:285-338.

[13]陈伟,熊波.人工智能刑事风险的治理逻辑与刑法转向:基于人工智能犯罪与网络犯罪的类型差异[J].学术界,2018(9):74-91.

[14]蔡曙山,薛小迪.人工智能与人类智能:从认知科学五个层级的理论看人机大战[J].北京大学学报(哲学社会科学版),2016,53(4):145-154.

[15]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.

[16]李彦宏.智能革命:迎接人工智能时代的社会、经济与文化变革[M].北京:中信出版集团,2017:122.

[17]翟中东.刑罚问题的社会学思考:方法及运用[M].北京:法律出版社,2010:122.

[18]萨宾娜·格莱斯,艾米丽·西尔弗曼,托马斯·魏根特.若机器人致害,谁将担责?[M]//陈世伟,译.陈兴良.刑事法评论(第40卷).北京:北京大学出版社,2018.

[19]马荣春.刑法公众认同研究[M].北京:中国政法大学出版社,2015:123.

[20]SCHERER M U. Regulating artificial intelligence systems: Risks, challenges, competencies, and strategies[J]. Harvard Journal of Law & Technology,2016,29:353-400.

[21]DURKHEIM E. The Rules of Sociological Method[M]. New York:Free Press,1938:23.

[22]刘宪权.人工智能时代刑事责任与刑罚体系的重构[J].政治与法律,2018(3):89-99.

[23]雷·库兹韦尔.人工智能的未来:揭示人类思维的奥妙[M].盛杨燕,译.杭州:浙江人民出版社,2016:6.

[24]孙道萃.人工智能对传统刑法的挑战[N].检察日报,2017-10-22(03).
[25]简爱.一个标签理论的现实化进路:刑法谦抑性的司法适用[J].法制与社会发展,2017,23(3):22-35.
[26]张明楷.外国刑法纲要[M].北京:清华大学出版社,2016.
[27]约翰内斯·韦塞尔斯.德国刑法总论[M].李昌珂,译.北京:法律出版社,2016:47.
[28]杜里奥·帕多瓦尼.意大利刑法学原理[M].陈忠林,译.北京:中国人民大学出版社,2009:229.
[29]SEMMLER S, ROSE Z. Artificial intelligence: Application today and implications tomorrow[J]. Duke Law & Technology Review, 2018,16:85-99.
[30]张青波.自我规制的规制:应对科技风险的法理与法制[J].华东政法大学学报,2018,21(1):98-111.
[31]ASIMOV I. I, Robot[M]. New York: Spectra,2008:5.
[32]SCHULLER A L. At the crossroads of control: The intersection of artificial intelligence in autonomous weapon systems with international humanitarian law[J]. Harvard National Security Journal,2017,8:379-425.
[33]白建军.中国民众刑法偏好研究[J].中国社会科学,2017(1):143-163,207-208.
[34]艾志强,沈元军.科技风险与公众认知的关系研究[J].中国人民大学学报,2012,26(4):107-114.
[35]陈兴良.本体刑法学[M].3版.北京:中国人民大学出版社,2017:236.
[36]黄荣坚.基础刑法学[M].3版.北京:中国人民大学出版社,2009:75.
[37]王政勋.刑法解释的语言论研究[M].北京:商务印书馆,2016:77.
[38]陈伟,熊波.利用信息网络犯罪行为二元形态的教义解读[J].上海财经大学学报(哲学社会科学版),2018,20(2):125-138,152.

On the system orientation and legislative attribute of artificial intelligence criminal risk

XIONG Bo

(*Law School, Southwest University of Political Science and Law, Chongqing 401120, P. R. China*)

Abstract: Artificial intelligence criminal risk does not belong to a transpersonal risk type. By questioning the subjective illusion of AI's criminal risk perception one by one, we can know that transpersonal risk can be divided into phenomenon risk at the factual level and the legal risk at the normative level. Beyond the scope of design and programming, the serious social harmful behavior of intelligent products is only a kind of pure fact phenomenon risk. What blocks the evaluation of criminal liability of AI products is the lack of experience summary of life emotion in AI technology itself. Applying penalty norms to AI products lacks practical significance, and in-depth learning relies on the basic output of human thinking mode. Legislative imputation of AI criminal risk should establish the subordinate nature of restriction, that is, to limit the existence of AI risk of permissibility and transcendence, and to clarify that AI criminal risk belongs to the subject of natural person. Then, it can lay a theoretical foundation for the scientific exploration of criminal law legislation in the era of artificial intelligence.

Key words: artificial intelligence; transpersonal risk; intelligent products; criminal liability

(责任编辑 胡志平)