

论人工智能安全管理义务 及其承担者的刑事责任

皮 勇^{*}

内容提要 对于人工智能系统自主决策和控制造成的危害后果,现行法律规制存有空白。若仅采取民事、行政法律手段,不足以遏制此类严重危害,需要引入刑法治理。关于刑法治理的路径,人工智能系统犯罪主体说难以实行,而妨害人工智能安全管理秩序犯罪立法的路径能够兼顾人工智能安全与发展,其中,采取违反人工智能安全管理义务犯罪的立法方案能够有效遏制前述危害。人工智能安全管理义务是为防控人工智能安全风险而设定的特殊、有限的规范责任。为有效惩治严重违反该义务的行为,有必要在刑法中增设违反人工智能安全管理义务罪。在罪行构造上,应依据义务犯、风险分配等刑法学理论,以有效降低人工智能安全风险为标准,合理规定该罪的人工智能安全管理刑法义务、主体和罪过等要件。

关键词 人工智能安全 风险 安全管理义务 义务犯 规范责任

DOI:10.14111/j.cnki.zgfx.2026.02.006

目 次

- 一、人工智能危害治理的法律路径
- 二、人工智能安全管理义务的性质与类型
- 三、人工智能安全管理义务承担者的刑事责任

^{*} 同济大学法学院教授。本文系 2021 年度国家社科基金重点项目“新发展理念与总体国家安全观视域下我国网络安全的刑法保障体系研究”(项目批准号:21AZD082)的阶段性成果。

随着人工智能的广泛应用,社会各领域都存在人工智能安全风险。在人工智能系统自主决策、控制的情况下,作为违法犯罪构成条件的主体、主观要件“消失”,依照现行法律没有人应对危害后果承担法律责任,导致人工智能安全风险出现制度性失控,以规范人类活动为潜在信条和被认为具有促进人类伦理互动价值观的现行法律,在此面临根本性的规制失灵。人工智能安全关系国家安全、社会公共安全和公众合法权益,需要在兼顾促进人工智能快速发展应用和保障安全的基础上,为人工智能安全发展提供充分的法律保障,包括刑法保障。

一、人工智能危害治理的法律路径

人工智能危害表现为多种类型:第一类是利用人工智能技术实施违法犯罪,如利用智能飞行器实施偷拍、伤害甚至杀人;第二类是人工智能系统研发应用过程中侵犯数据安全和个人信息权益等的违法犯罪;第三类是人工智能系统自主决策、控制造成危害后果,如生成式人工智能传播虚假有害信息、完全自动驾驶汽车造成人员伤亡等。我国现行刑法可以适用于前两类危害,对第三类危害则“力有不逮”。

(一) 人工智能危害的法律治理主张

人工智能危害的刑法治理是人工智能危害法律治理体系的重要组成部分,二者在治理策略和逻辑上应当保持一致,惟此才能构建各部门法相互衔接配合的、系统化的法律治理体系。关于前述第三类危害的法律治理路径,主要有人类代理说、产品责任说、拟制主体说、安全管理义务说四种主张。

1. 人类代理说

人类代理说认为,人工智能系统的功能和决策方式可以直接追溯到系统的设计、编程和知识,人定义、引导和最终控制着系统的研发和应用,无论系统多么复杂,其本质上是人类使用的工具,不具有法律人格,应当作为人的代理,由自然人或其集合体对人工智能系统引起的危害承担责任。^①例如,在“达芬奇”等手术机器人自主手术造成病患伤害的案件中,应追究机器人生产厂家、医院而不是手术机器人的责任。

这种观点没有考虑人工智能系统自主运行的事实。确有人工智能系统的运行没有遵照人类的指令、受人类控制或在人类控制能力范围内的情形,如L5级自动驾驶汽车完全自主行动,其自主决策的场景中有的是研发设计和应用人员无法预见的,更谈不上控制。完全自主的人工智能系统处于没有人类参与或干预的独立“感知—分析—行动”的“行为状态”,“能够独立主动地制定自己的计划”,有学者认为其更适合被视为人而不是机器。^②笔者认为,人工智能的自主性动摇了人类代理说的事实基础。除了代理人

^① See Ryan Abbott & Alex Sarch, *Punishing Artificial Intelligence: Legal Fiction or Science Fiction*, UC Davis Law Review, Vol.53:323, p.369-373 (2019).

^② See Nick Bostrom, *When Machines Outsmart Humans*, Futures, Vol.35:759, p.763 (2003).

必须是自然人或其集合体,一般情况下代理人应当按委托人的指示行事,表见代理和紧急情况下例外属于特别情形,而人工智能系统的完全自主活动并不限于前述特殊情形,是一种常态,不满足代理的条件,将危害后果归责于其使用者或生产者,缺乏事实和法律依据。

2. 新产品责任说

新产品责任说主张构建人工智能产品的新严格责任制度,生产者应当对教会系统“思考”的附生后果(即造成不可预见的危害)承担严格责任。^③人工智能系统自主行动引起的危害应归因于产品的制造缺陷、设计缺陷或编程不佳等。可是,人工智能系统更加复杂,即使是按“消费者期望”(consumer expectations)和“风险效用”(risk-utility)标准,证明存在以上缺陷也是困难的,导致受害人因无能力证明而得不到赔偿。对此,该观点认为,如果能够排除人为责任(如驾驶位人员或安全监督员的过错责任),即使不能查明人工智能系统存在缺陷,当事故的发生有一定频率且表现出共同特征时,应当认定为产品缺陷。此外,人工智能系统由多方提供技术和设备,如无人驾驶汽车的自动驾驶系统、雷达和激光传感器等,这些软件和硬件设备容易出现无法检测的故障,特别是人工智能算法系统具有不可解释、不可预测性。只追究作为人工智能系统生产者的责任是不公平的,应将人工智能系统关联公司按照“共同企业”对待,追究其中部分公司的连带责任,以解决无法将故障归因于其中特定公司或自然人的追责困难。^④

持相似的新严格责任观点的学者认为,事实因果关系和法律因果关系检测遇到人工智能暗箱障碍,应通过对人工智能的创造者施加较高标准的注意义务,对违反该义务、威胁公共安全的行为,按严格责任原则追究法律责任,目的是“允许超越因果关系检验,将有关的损害与创造者的过失行为联系起来”,以平衡人工智能发展和安全的需要。^⑤

笔者认为,该观点认识到人工智能系统的自主性和复杂性,以及按照传统产品责任制度追责的缺陷,但存在以下不足:其一,对产品故障或缺陷进行推断认定,并延展责任追究的因果链条,要求人工智能系统生产者及为其提供技术或零部件的上游企业承担连带责任的观点,实际上是要求这些单位在未被证明有过错的情况下承担补偿义务,既脱离了现行产品责任和侵权责任的法治轨道,也没有解决相关企业之间的责任分配问题。其二,只对人工智能系统的生产者施加注意义务及按严格责任制度追究法律责任,不能降低人工智能系统交付后因其自主学习产生的安全风险,没有考虑到使用者直接监督系统运行的作用。其三,人工智能的自主性导致其不可预测,生产者对危害后果缺乏预见和预防的可能性,而依照严格责任制度及其理论,只有人工智能系统的自主决策和活动

^③ See Kyle Colonna, *Autonomous Cars and Tort Liability*, Journal of Law, Technology & the Internet, Vol.4:81, p.109, 116-117 (2012).

^④ See David C. Vladeck, *Machines without Principals: Liability Rules and Artificial Intelligence*, Washington Law Review, Vol.89:117, p.138-140, 148-150 (2014).

^⑤ See Zhao Yan Lee, Mohammad Ershadul Karim & Kevin Ngui, *Deep Learning Artificial Intelligence and the Law of Causation: Application, Challenges and Solutions*, Information & Communications Technology Law, Vol.30:255, p.278-280 (2021).

是可预测的,让生产者承担严格责任才是公平的。^⑥其四,受人工智能系统侵害的不仅有被害人的财产性利益,还可能有生命、健康和其他重要权益,仅让生产者及相关企业进行民事赔偿或补偿,不利于保护公众生命、健康等重要法益。这在历史上是有教训的,在巨大的产业利益诱惑下,单纯地依靠民事赔偿不足以有效控制安全风险。^⑦

3. 拟制主体说

拟制主体说认为,生产者无法预见人工智能系统的全部“行为”,应承认后者具有法律能力(legal capacity),因其在功能上能够“有目的地行事”,展现出所谓“道德的”“法律的”能力。人们像对待自然人一样对待拟人化的机器人,应赋予后者权利使其免受人类的虐待。拉丁语“人格”(persona)的原意就是面具或角色,既然可以承认各种非自然人主体(如公司)的法律人格,对比公司法人更“实在”的人工智能系统也可以赋予其法律人格并制定制裁措施。^⑧持该种观点的学者认为,人工智能系统可以被追究民事和刑事责任,并建议增设删除系统软件、限制自由、社区服务和罚款等专门的刑罚。^⑨还有学者将人工智能系统的法律拟制人格区分为伦理人格和技术人格,认为人工智能的人格将“以技术人格的探索先行,逐渐进行伦理人格的塑造,人类或机器人的伦理人格最终成为技术人格的依归”。^⑩拟制人格说曾得到国外立法机构的响应,2017年欧洲议会提交《关于机器人的民法规则的决议》,呼吁欧盟委员会在人工智能立法中,“为机器人创造一个特定的法律地位,至少最复杂的自主机器人可以被确定为具有电子人地位,使其可以负责修复可能造成的任何损害,并将电子人格应用于机器人自主决定或与第三方独立互动的情形”。^⑪

笔者认为,该种观点存在以下不足:其一,人工智能系统与公司法人不具有可比性,前者不具有独立于自然人或法人的独立利益。其二,未来将有数以亿计的自主人工智能系统投入应用,授予其以主体资格的现实性和可行性存在疑问。即使可行,能否消除人工智能危害及其效率同样存疑,反而会因免除或弱化现有法律主体防控安全风险的责任,导致“有组织的不负责任”^⑫。其三,采纳该观点则要对现行法律制度作重大修改,需要制定适用于“第三类法律主体”的法律制度,法律变革成本大,效益不高,难以与现行

⑥ See Mohammad Bashayreh, Fadi N. Sibai & Amer Tabbara, *Artificial Intelligence and Legal liability: Towards an International Approach of Proportional Liability Based on Risk Sharing*, *Information & Communications Technology Law*, Vol.30:169, p.172-174 (2021).

⑦ See Frank J. Vandall, *Criminal Prosecution of Corporations for Defective Products*, *International Legal Practitioner*, Vol.12:66, p.67-68 (1987).

⑧ 参见[德]霍斯特·艾丹米勒:《机器人的崛起与人类的法律》,李飞、敦小匣译,载《法治现代化研究》2017年第4期,第68-69页。

⑨ See Gabriel Hallevy, *The Criminal Liability of Artificial Intelligence Entities - From Science Fiction to Legal Social Control*, *Akron Intellectual Property Journal*, Vol.4:171, p.194-199 (2010).

⑩ 参见梅夏英:《伦理人格与技术人格:人工智能法律主体地位的理论框架》,载《中外法学》2025年第1期,第26、41-43页。

⑪ See Civil Law Rules on Robotic, OJ C 252, 18.7, point 59(f), <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52017IP0051>.

⑫ 叶良芳、马路瑶:《风险社会视阈下人工智能犯罪的刑法应对》,载《浙江学刊》2018年第6期,第67-68页。

法律制度保持协调。对人工智能系统追究刑事责任,不能实现刑法的目的与任务,与犯罪、刑事责任和刑罚基本制度不相容。

4. 安全管理义务说

安全管理义务说认为,人工智能系统是自主性机器和人类的工具,人工智能系统安全管理义务承担者应确保其安全并对其危害后果负责。美国《人工智能倡议法》和欧盟《人工智能法》都将人工智能系统定义为“基于机器的系统”,要求“人工智能系统被开发和用作一种服务于人类、尊重人类尊严和个人自主权的工具,并以一种可以被人类适当控制和监督的方式运作”(欧盟《人工智能法》引言第27条)。^⑬后者还制定了人工智能安全风险防控制度,要求人工智能系统提供者、部署者^⑭及其他相关方承担安全管理义务和相关法律责任。以上法律既未赋予人工智能系统法律人格,也未规定其法律责任。我国人工智能发展战略文件及相关立法持相同立场,我国提出的《全球人工智能治理倡议》要求“确保人工智能始终处于人类控制之下,打造可审核、可监督、可追溯、可信赖的人工智能技术”,《互联网信息服务深度合成管理规定》等将人工智能系统作为防控对象,要求人工智能系统提供者及其他相关方承担安全风险管理义务。该观点客观反映了人工智能的特性,能有效防控人工智能危害,符合促进人工智能安全发展的国际共识,对现有法律体系的冲击不大,对人工智能危害的法律治理具有重要的指导作用。

以上四种观点存在的共同不足在于,主要以民事、行政手段来治理人工智能危害,解决被害人的赔偿或补偿问题,不能遏制严重危害。事实上,人工智能危害不仅威胁公众的生命、健康和财产安全,还严重危害社会公共安全和国家安全,民事、行政法律手段不足以遏制严重的人工智能危害,有必要为人工智能安全提供刑法保障。

(二) 人工智能危害治理的刑法路径

对于人工智能危害的刑法治理,前述第三类人工智能危害客观上不能归因于自然人的行为,主观上没有人对其有罪过,不能采用人类代理说。由于主客观相统一原则是我国刑法的基本原则,故也不能采纳产品严格责任说。目前以下两种观点较有影响。

1. 人工智能系统犯罪主体说

该观点认为,人工智能系统自主决策并直接引起危害后果,可以将其拟制为犯罪主体,并依照现行刑法追究刑事责任,^⑮这符合目的刑和报应刑的立场。通过对人工智能系统定罪和处以特定刑罚,如限制应用、删除数据、修改或销毁系统等,可以剥夺人工智能系统的再犯能力,威慑其他人工智能系统,还可以促使其所有者制止危害,避免其系统被处罚而遭受损失。对人工智能系统处以与其危害相称的刑罚,支持被害人的权益,可以安抚被害人感情,表达官方的谴责,恢复公平正义和社会安全感。^⑯

^⑬ See Artificial Intelligence Act, (EU) 2024/1689, <http://data.europa.eu/eli/reg/2024/1689/oj>.

^⑭ 欧盟《人工智能法》第3条中的提供者(Provider)和部署者(Deployer)与我国相关规章中使用的概念有较大差别,后者使用的“服务提供者”相当于前者规定的部署者。

^⑮ 参见刘宪权:《ChatGPT等生成式人工智能的刑事责任问题研究》,载《现代法学》2023年第4期,第115-121页。

^⑯ See Jack Boeglin & Zachary Shapiro, *A Theory of Differential Punishment*, *Vanderbilt Law Review*, Vol.70:1499, p.1503-1504 (2017).

该观点实际上难以实施,首先有主体资格障碍。刑法只处罚有罪责能力的行为人,罪责能力是刑法适用的必要条件,人工智能系统不具有刑法上认识和意志的生物学和社会学基础,既没有意识和感觉,也没有利益或幸福,不满足犯罪主观要件的基础,不具有受刑法处罚的资格。为了解决这一问题,学者们尝试了多种理论工具,主要有:

(1) 雇主责任理论。该观点认为,公司是拟制的犯罪人,本身不具有受谴责的能力,^{①7}借助于雇主责任理论,刑法赋予公司可谴责的主观状态。公司雇员在受雇范围内行事且服务于公司利益,雇员的主观状态被认为是公司的主观状态,^{①8}对公司定罪处刑不违反法治原则。按此理论,人工智能系统也可以被赋予罪过状态和罪责能力。笔者认为,人工智能系统自主决策与公司决策的机制不同,系统的提供者、使用者、所有者对危害没有罪过,何以作为人工智能系统的罪过状态和罪责能力的依据?该理论不能解决人工智能系统罪责能力问题。

(2) 客观责任理论。该观点认为,从结果主义的立场出发,缺乏有罪的精神状态也可以定罪,“可能发展出一种合理的客观标准作为人工智能罪责能力的判断依据”,^{①9}没有罪过心态的人工智能系统也可追究刑事责任。笔者认为,即使是严格责任犯罪,作为刑法意义上的行为也必须基于精神状态,应当维护这一现代刑法制度的基础。行为首先是能够归于作为心理和精神的动作中心的自然人的一切,是“人格表现”。^{②0}人工智能系统没有精神状态,其行动不能被评价为刑法上的行为,否则将颠覆刑法谴责有罪过行为的理论基础。

(3) 人工智能系统的特别犯罪意图理论。该观点认为,罪责性是有罪的心理以不尊重法益的方式表现出来,即以行为表现,而不是动机、思想和感受方面的细微差别,来证明有罪的心理状态。^{②1}法律允许以公司员工收集信息和决策的行为,认定公司存在不尊重法益的罪责心理,那么,也可以采取类似的方式来认定人工智能系统的罪责心理。例如,自动驾驶系统检测到道路上的运动目标,如果其控制汽车驶向目标,而不是减速或避开,可以依据检测数据和控制信息等,证明该汽车有促成危害结果发生的罪过心理。自然人对某事的真实性有充分稳定、强烈的倾向时,即可认为有认识,那么,人工智能系统有同样的倾向时,也应认定系统有认识。该观点主张按照以上规则来认定人工智能系统的罪过心理,并制定认定人工智能系统犯罪的司法规则。^{②2}笔者认为,该理论存在转移问题的嫌疑,将待证明的结果作为证明的前提,是先假定人工智能系统存在罪过心理,而后提出认定罪过的规则,并没有解决罪过的存在问题。

^{①7} See Albert W. Alschuler, *Two Ways to Think about the Punishment of Corporations*, *American Criminal Law Review*, Vol.46:1359, p.1367, 1369 (2009).

^{①8} See Ryan Abbott & Alex Sarch, *supra* note 1, 351.

^{①9} 参见江溯:《人工智能作为刑事责任主体:基于刑法哲学的证立》,载《法制与社会发展》2021年第3期,第126页。

^{②0} 参见[德]克劳斯·罗克辛:《德国刑法学总论》(第1卷),王世洲译,法律出版社2005年版,第160页。

^{②1} See Alexander Sarch, *Who Cares What You Think? Criminal Culpability and the Irrelevance of Unmanifested Mental States*, *Law and Philosophy*, Vol.36:707, p.707-709 (2017).

^{②2} See Ryan Abbott & Alex Sarch, *supra* note 1, 357-360.

其次,人工智能犯罪主体说的实施有可罚性障碍。刑罚的本质是“给受罚人以痛苦的制裁措施”^{②③}。这里的“痛苦”应当理解为使受刑者感受到痛苦,或其他通常被认为是不愉快的后果,后者包括一般人认为是不愉快但并未使受刑者不愉快的后果。这两种形式的痛苦都缘于对受刑者利益的限制或剥夺,即使受刑者不因其利益被限制或剥夺而感到痛苦,客观上仍然对其造成不好的后果。人工智能系统不能感到痛苦或不愉快,尽管丧失能力或被摧毁在客观上剥夺系统的利益,可以认为是对系统的刑罚,但利益由欲望和目标组成,二者以看法和意识为前提,受刑者至少要有快乐和痛苦的体验,才能真正拥有利益。人工智能系统不具有真正的认知和意识能力,不具有纯粹意义的利益,对其利益的剥夺也就谈不上是刑罚。

也有观点主张将定罪和处刑分开,处刑不作为定罪的前提。例如,对公司不能处以生命刑、自由刑等,不影响对公司定罪,刑法规定由主管人员和直接责任人替代公司受罚。对人工智能系统定罪后,也可以由其提供者、使用者、所有者替代受罚。即使人工智能系统不具有可罚性,从结果主义的立场看,对其定罪也有意义,能够表明刑法对人工智能犯罪的严厉谴责。^{②④}笔者认为,刑法规定对单位犯罪的刑罚中,对公司主管人员和直接责任人员的处罚,不应被定性为替代公司受罚,以上人员与公司的关系也不同于人工智能系统与其提供者、使用者、所有者的关系,因为前者的行为和心态产生公司的意志,而后者则无此决定能力。而且,无刑事责任的定罪违反罪责刑相适应原则,不能实现维护公正和预防犯罪的目的。

2. 妨害人工智能安全管理秩序犯罪说

人工智能系统是新的工具和治理对象,人工智能安全是人工智能时代的新型公共安全。为了保护人工智能安全,我国正在制定人工智能安全管理制度,建立人工智能安全管理秩序,要求相关人员遵守人工智能安全行为规范,防控人工智能安全风险。为了惩治严重侵犯人工智能安全的行为,我国应当设立妨害人工智能安全管理秩序犯罪,对严重违反前述行为规范的行为追究刑事责任。妨害人工智能安全管理秩序犯罪包括滥用人工智能犯罪和违反人工智能安全管理义务犯罪。前者不同于传统犯罪的智能化,侵犯的不是传统法益而是新型利益,如滥用人工智能操纵自然人以损害其合法权益、对自然人进行生物识别分类、不公正的社会评分、风险分析、人脸采集、情绪分析等行为,不构成现行刑法中的犯罪,却严重损害公众的自由、生活、工作等利益。后者是为了防控人工智能安全风险而设置的犯罪。人工智能系统具有自主、智能特性,传统的产品安全保障制度不能控制其安全风险。人工智能安全风险源于研发并沿着其价值实现链扩散,人工智能研发、生产、提供、部署、终端使用各方对风险形成、扩散有原因力,且具有直接管控风险的能力,应承担安全管理义务。设立义务犯罪可以督促相关人员切实履行前述义务,防控严重风险,保障人工智能安全。人工智能安全与网络安全、数据安全是具有相似信

^{②③} [意] 杜里奥·帕多瓦尼:《意大利刑法学原理》,陈忠林译,法律出版社1998年版,第342页。

^{②④} See Ryan Abbott & Alex Sarch, *supra* note 1, 367.

息技术特性的新型集体法益,我国《刑法》将危害网络安全的犯罪规定在“妨害社会管理秩序罪”一章的“扰乱公共秩序罪”一节中,设立了拒不履行信息网络安全管理义务罪等妨害信息网络安全管理秩序犯罪,妨害人工智能安全管理秩序犯罪侵犯的是同类客体,都是现代社会环境下的新型公共秩序。

该主张具有合理性。首先,该主张与国际社会治理人工智能安全的方向一致。《全球人工智能治理倡议》要求“明确人工智能相关主体的责任和权力边界”,“确保人工智能始终处于人类控制之下”。欧盟《人工智能法》将人工智能系统定性为自主性工具,建立了类型化、全过程、多方协作的安全风险防控制度。^⑤妨害人工智能安全管理秩序犯罪立法的立场和路径与前二者一致。其次,该主张客观反映了人工智能危害的形成与防控规律。无论是采取何种技术方法、何种自主性程度的人工智能系统,提供者和部署者对于系统导致的危害或风险有引起或扩散作用,如果具备风险管理能力,理应承担防控责任,设置以上犯罪能够督促其履行防控义务,有效降低安全风险,符合人工智能危害的形成与防控规律。最后,该主张能够兼顾保护法益和保障人工智能发展。该立法主张依据风险防控理论,兼顾人工智能安全与发展,既能有效降低安全风险,又按风险高低分类设置行为规范,采取基于风险的差别化刑法应对,能够最大限度地保障人工智能发展。

在妨害人工智能安全管理秩序犯罪中,滥用人工智能系统犯罪和违反人工智能安全管理义务犯罪是两类不同的犯罪,后者可以遏制第三类人工智能危害,是本文主要研究的内容。

二、人工智能安全管理义务的性质与类型

人工智能安全相关人员履行安全管理义务能够有效降低风险,违反人工智能安全管理义务的犯罪立法能够惩治严重违反安全管理义务的行为,为人工智能安全管理制度提供刑法保障。人工智能安全管理义务是该类犯罪的构成要件要素,也是国内外人工智能安全立法的重要内容,有必要研究其性质与类型。

(一) 人工智能安全管理义务的性质

设置人工智能安全管理义务是为了防控安全风险,有必要明确人工智能安全风险的内涵。风险是指“与未来损失有关的不确定性”,预防风险就是要降低损失发生的可能性或损失的程度。^⑥欧盟《人工智能法》第3条第2款将风险定义为“发生危害的可能性和危害的严重性的组合”,与危害发生的概率和危害的严重性正相关。风险是可能发生的危害,如果某活动能够确定造成危害后果,其引起的不是风险而是确定的危害,应当被制止。风险意味着蕴含利益的机会,只能采取措施控制风险,而不应禁止有风险的活动,否则会失去获取利益的机会。人工智能安全风险可以定义为人工智能研发、生产、提

^⑤ 参见皮勇:《欧盟〈人工智能法〉中的风险防控机制及对我国的镜鉴》,载《比较法研究》2024年第4期,第69-80页。

^⑥ 参见〔德〕尼克拉斯·卢曼:《风险社会学》,孙一洲译,广西人民出版社2020年版,第42、51页。

供、应用过程中发生危害的可能性和危害的严重性的组合,具有全领域渗透性、不可预见性、网络传播性等特点,能够在社会各领域造成难以有效控制的严重后果。防控传统风险的法律制度因不适应其特点而难以有效控制,不能为社会公共安全和公众合法权益提供充分保障,有必要建立新的风险防控制度。

现行刑法不能规制前述第三类危害的原因,是相关人员的行为与危害后果之间不存在刑法上的因果关系,没有因果关系就无法适用现行刑法规定的犯罪。由于人工智能系统自主、直接引起危害后果,故其研发、生产、提供和使用方的行为与危害后果之间没有因果关系。以机器学习算法为例,该算法基于训练数据中的相关性,如 ChatGPT 的算法逻辑是基于数以亿计的参数,对海量训练数据加权统计后,形成复杂的关联关系。^{②7}从最初的算法设计者到终端用户,各方行为对算法的生成和演进不起决定性作用,不能评价为原因行为。由于欠缺因果关系,依照现行法律不能要求前述人员对危害后果承担法律责任。如果将因果关系泛化为相关性关系,则与条件说无异,实际上是“舍弃因果关系要件设定针对违法行为本身的责任,以此跳过因果联结及可追溯性难题”^{②8},将导致人工智能安全风险防控主体范围的过度扩展,动摇现行法律责任制度的理论基础。因此,“以人的活动为中心”的社会环境下发展起来的因果关系理论,在智慧社会环境下无法为解决人工智能危害的法律治理困难提供理论指导。只有改变按结果犯、支配犯治理的思维习惯,在人工智能安全治理领域完成基于风险的、义务导向的归责模式转向,^{②9}才能摆脱当前治理困境。

对人工智能安全相关人员施加安全管理义务,在当前情形下是治理前述第三类人工智能危害唯一可行的选择,规范责任理论可以为其提供理论依据。从现实情况来看,要有效降低人工智能安全风险,只能通过人工智能安全相关人员履行安全管理义务来实现。从法律规范看,设立安全管理义务是人工智能安全保障体系发展的必然结果。为了保障人工智能安全,国家必然要制定人工智能安全管理制度,设定人工智能安全相关人员的行为规范。人工智能安全管理义务是对其设定的特殊规范责任,对有能力遵守而故意违反行为规范的行为人追究法律责任。刑法学中的规范责任理论可以为以上法律制度提供理论支持,其主张承担刑事责任的根据是“违反义务实施违法行为的决意”,包括责任能力、故意或过失以及“行为人存在适法行为的期待可能性”,^{③0}违反规范义务就应承担相应的法律责任。

不过,人工智能安全管理义务不同于传统的规范责任,是有限的规范责任。其原因主要是两个方面:其一,因人工智能系统自主性引起的危害后果与前述人员行为之间只有非直接的、弱相关性关系,而不是因果关系,依照现行法律其不应对危害后果承担法律

^{②7} 参见张熙、杨小汕、徐常胜:《ChatGPT 及生成式人工智能现状及未来发展方向》,载《中国科学基金》2023 年第 5 期,第 744-745 页。

^{②8} 王莹:《算法侵害责任框架刍议》,载《中国法学》2022 年第 3 期,第 174-176 页。

^{②9} 参见陈璇:《社会治理视角下刑法归责模式的现代转型》,载《中国社会科学》2025 年第 7 期,第 114-121 页。

^{③0} 参见马克昌:《比较刑法原理:外国刑法学总论》,武汉大学出版社 2002 年版,第 437 页。

责任。从义务规范角度对义务承担者施加的责任,也应与其影响风险的作用和防控能力相适应,较之于传统的规范责任要有所缩减,以区别于传统义务犯的义务。其二,人工智能安全风险沿着价值链扩展,从系统提供者、使用者到终端用户,各方只面对其所在阶段上的风险,按照风险与注意义务分配的对应关系,只应对部分风险承担防控责任。^{③①}对于应追究法律责任的危险或实害而言,各方承担的风险防控责任,既要与其所在阶段上的风险大小成比例,也要与其地位和防控能力成比例,否则违反比例原则中的均衡性原则。^{③②}因此,人工智能安全管理义务应当是为防控风险所必要的、适度且有限的规范责任,与相关人员引起或扩散风险所起的作用或防控风险的能力成正比。

人工智能安全管理义务与信息网络安全管理义务都是安全管理义务,但二者有本质区别:(1)法律性质不同。前者是对“物的安全”的管理义务,防控作为工具的人工智能系统自主运行引起的安全风险,近似于生产作业安全管理义务。后者是针对“人的活动”的监督管理义务,防止他人实施违法犯罪。有论者在讨论生成式人工智能产品提供者的安全管理义务时,将该义务限于提供者对他人利用其产品进行违法犯罪活动的管理义务,即后者义务。^{③③}生成式人工智能服务提供者同时也是网络信息服务的提供者,当然应承担信息网络安全管理义务,但是,该义务不同于人工智能安全管理义务。(2)义务根据不同。后者的设立根据是网络服务提供者的保证人地位,即“保护他人法益不受来自于自己控制领域的危险威胁的义务。这种对于危险源的控制是不作为犯的义务”^{③④}。前者的设立根据是人工智能安全相关人员引起或扩散人工智能安全风险的客观事实和保护社会安全的需要,以上人员并不具有保证人地位。(3)保护法益不同。前者直接保护的是人工智能安全,间接保护的法益范围广泛,如公共安全和公众合法权益等,而后者保护的是公共信息网络安全管理秩序。^{③⑤}由于以上区别,不对违反前者义务的行为适用拒不履行信息网络安全管理义务罪。^{③⑥}

(二) 人工智能安全管理义务的类型

国内外人工智能安全立法都规定了人工智能安全管理义务。欧盟《人工智能法》规定了基于风险的人工智能安全管理义务制度,针对人工智能系统和通用人工智能模型两类防控对象,将义务承担者分为提供者、部署者、欧盟内授权代表、进口者、经销者等,从防控对象和义务承担者两个维度规定了不同的义务类型。^{③⑦}美国《人工智能倡议法》提出了发展“可信的人工智能系统”,算法应满足“可解释性”“能够识别和缓解人工智

^{③①} See Mohammad Bashayreh, Fadi N. Sibai & Amer Tabbara, *supra* note 6, 180-185.

^{③②} 参见谢立斌:《比例原则的适用范围与审查基准》,载《当代法学》2024年第5期,第87-89、95页。

^{③③} 参见刘杰:《ChatGPT类生成式人工智能产品提供者之刑事责任》,载《法治研究》2024年第2期,第63-66页。

^{③④} [德] 乌尔里希·齐白:《网络服务提供者的刑事责任——刑法总论中的核心问题》,王华伟译,载赵秉志主编:《刑法论丛》(第48卷),法律出版社2016年版,第330页。

^{③⑤} 《刑法》将拒不履行信息网络安全管理义务罪设置在“扰乱公共秩序罪”一节中,其保护法益应当是公共信息网络安全管理秩序。

^{③⑥} 参见刘宪权:《涉生成式人工智能数据犯罪刑法规制新路径》,载《当代法学》2024年第6期,第12-13页。

^{③⑦} 参见前注^{②⑤},皮勇文,第72-78页。

能系统中偏见的分析方法”“具有安全、稳健性”等要求,^⑳美国联邦政府还发布了《推进国家安全领域人工智能治理和风险管理的框架》,规定了基于风险的人工智能应用分类管理制度。近年来我国制定了《互联网信息服务深度合成管理规定》《生成式人工智能服务管理暂行办法》《人脸识别技术应用安全管理规定》《人工智能生成合成内容标识办法》等规章,规定了人工智能生成内容信息和人脸识别技术应用相关安全管理义务。对比以上规章,欧盟《人工智能法》规定了全面、完整的安全管理义务体系,我国的规定与之相似。以下主要将欧盟《人工智能法》规定的义务类型与我国相关规定进行比较。

1. 人工智能系统安全管理义务

欧盟《人工智能法》将人工智能系统分为高风险人工智能系统、非高风险的特定人工智能系统和其他人工智能系统,对前两种系统的相关人员规定了不同的安全管理义务,鼓励和促进制定针对第三类人工智能系统的行为守则,并提倡相关人员自愿遵守该法的要求(第95条)。

(1) 高风险人工智能系统安全管理义务

高风险人工智能系统是指对人的健康、安全或基本权利造成伤害的风险高,且用于该法规定领域的人工智能系统(第6条第1—3款),是欧盟《人工智能法》的主要防控对象。该类系统的提供者、部署者及其他方被要求承担较多的安全管理义务。

为了确保高风险人工智能系统投放市场或投入使用不会对欧盟法律所认可和保护的公共利益构成不可接受的风险(引言第46条),欧盟《人工智能法》要求高风险人工智能系统符合以下7项要求:(1)建立实施、记录和维护风险管理系统;(2)数据治理;(3)提供技术文件;(4)自动记录日志;(5)保持透明度;(6)适合人工监督;(7)保持系统的准确性、鲁棒性和网络安全(第8—15条)。此类系统提供者的义务是确保高风险人工智能系统符合这些要求,此外,提供者还要承担以下6项义务:(8)遵守一般管理程序要求;(9)建立质量管理体系;(10)制作、保存和随时向国家主管机构提供该法第18条规定的文件;(11)保存日志数据;(12)对不符合要求的高风险人工智能系统采取必要的纠正措施;(13)与主管机构合作(第16—21条)。在以上义务中,对降低安全风险发挥实质性作用的是(1)(2)(5)(6)(7)(12)项义务,其它义务是监管程序性义务。在我国前述规章中,仅《人脸识别技术应用安全管理规定》中的人脸识别设备符合欧盟《人工智能法》规定的高风险人工智能系统。^㉑但该办法不适用于“为从事人脸识别技术研发、算法训练活动应用人脸识别技术处理人脸信息”的行为,没有规定人脸识别技术系统提供者的安全管理义务。如果系统提供者同时又是系统使用者,则适用该办法,其地位相当于欧盟《人工智能法》中的高风险人工智能系统的部署者。

高风险人工智能系统的安全风险既源于系统设计,也来自系统使用。部署者对系统使用环境、可能受影响的人或群体包括弱势群体有更准确的认识,更了解高风险人工智

^⑳ See U.S. Code, Title15, § 9461(7).

^㉑ 欧盟《人工智能法》附录三将用于8类领域活动的人工智能系统规定为高风险人工智能系统。

能系统的具体使用情况,能够发现开发阶段未预见的潜在重大风险,提供关于危险的信息,这对弥补提供者的义务缺漏、防控安全风险能够起到关键作用。^⑩ 欧盟《人工智能法》对部署者规定了以下12项义务:(1)遵守使用说明;(2)安排适格人员进行人工监督;(3)控制数据输入以符合预期目的;(4)监控系统运行;(5)保存自动生成的日志;(6)告知受影响的雇员;(7)注册登记;(8)评估数据保护状况;(9)对特定对象遵法使用系统;(10)告知系统的使用对象;(11)与相关国家主管机构合作等;(12)评估高风险人工智能系统可能对基本权利的影响(第26、27条)。其中,第(1)(2)(3)(4)(6)(9)(10)项义务是在应用阶段的、能够实质性降低安全风险的义务,与提供者在研发阶段的义务相互补充、配合。我国《人脸识别技术应用安全管理暂行办法》中,对应于欧盟《人工智能法》部署者义务的是告知和个人信息保护影响评估义务,^⑪ 没有规定专门的人工智能安全管理义务。人脸识别系统提供者的前述行为涉及敏感个人信息的处理,受《个人信息保护法》《民法典》等法律法规的规制。

高风险人工智能系统风险与人工智能价值链上多方行为紧密关联,欧盟《人工智能法》还对高风险人工智能系统进口者、经销者以及在欧盟境外的高风险人工智能系统提供者的授权代表规定了安全管理义务。这三类主体使高风险人工智能系统从提供者(包括境外提供者)流转 to 部署者,对风险扩散有一定作用,承担的主要是监督提供者和部署者、协助风险管理等监管程序性义务。我国没有类似的规定。

(2) 非高风险的特定人工智能系统安全管理义务

人工智能系统生成合成内容信息越来越逼真,其社会化应用增加了虚假信息、社会操纵、欺诈、假冒等新风险。为了防控此类风险,欧盟《人工智能法》要求非高风险的特定人工智能系统相关方承担透明度义务。一是与自然人直接互动的人工智能系统的透明度义务。该类系统提供者应确保系统在设计 and 开发时,能够使与系统互动的自然人知晓其面对的是人工智能系统;但根据当时的环境和使用情况,从合理的知情、观察和谨慎的自然人的角度看这一点是非常明显的除外。生成合成音频、图像、视频或文本内容的人工智能系统(包括通用人工智能系统)的提供者应当确保,人工智能系统的输出以机器可读的格式、且可检测为人工生成或操作的方式进行标记,该系统的部署者应当披露该内容是人工生成或操纵的。二是情绪识别系统或生物识别分类系统的透明度义务。该类系统的部署者应当告知自然人其是系统的识别对象,并按照欧盟相关法规和指令处理个人数据(第50条第1—4款)。如果以上系统符合高风险人工智能系统的特征,须遵守高风险人工智能系统相关义务。遵守以上透明度义务不应被解释为使用该类系统或其输出是合法的,相关方仍须遵守欧盟及成员国相关法律规定,如个人数据保护义务(第50条第6款)。

我国前述规章对互联网信息服务提供者和人脸识别设备使用者规定了类似的透明

^⑩ See Artificial Intelligence Act, C/2024/506, Recital (58a), <https://eur-lex.europa.eu/eli/C/2024/506/oj/>.

^⑪ 参见《人脸识别技术应用安全管理暂行办法》第5、9条。

度义务,^④ 对应于欧盟《人工智能法》中与自然人直接互动的人工智能系统部署者的透明度义务,但是在义务的内容、范围、义务主体等方面超过后者,如将义务主体扩大到提供网络信息内容传播服务的服务提供者、互联网应用程序分发平台以及网络用户,使下游环节的更多主体参与风险防控。

2. 通用人工智能模型安全管理义务

欧盟《人工智能法》对通用人工智能模型的提供者及其授权代表规定了安全管理义务。通用人工智能模型是人工智能系统的核心组成部分,处于下游的系统提供者只有充分了解模型,才能将其集成到系统中(引言第88条)。根据高影响能力标准(第51条),该法将通用人工智能模型分为有无系统性风险的两类模型,^⑤ 要求通用人工智能模型提供者承担提供技术文件、向下游人工智能系统提供者提供技术信息、制定合规政策、公开通用人工智能模型训练内容4项义务(第53条第1款),有系统性风险的通用人工智能模型的提供者还应承担模型评估、识别并减轻系统性风险、跟踪、记录和报告严重事件、网络安全保护4项义务(第55条第1款),欧盟境外设立的人工智能系统模型提供者的授权代表应承担制定、保存和提供技术文件与提供者的联系方式、同主管机构合作等义务(第54条第3款)。在以上义务中,只有识别和减轻系统性风险、网络安全保护义务属于实质性降低风险的义务,其他都属于监管程序性义务。不同于人工智能系统风险的现实性,通用人工智能模型在被提供给下游单位时,因为尚无具体的应用目标而不存在实际风险,其基于高影响能力的系统性风险需要通过系统应用才能被现实化,只是因其可能大量应用而产生重大负面影响,才需要提前采取防控措施。我国没有针对通用人工智能模型规定安全管理义务。

综上,人工智能安全管理义务是为了防控人工智能安全风险而对人工智能安全相关人员施加的有限的规范责任。我国仅对人工智能信息服务的提供者及相关方规定了安全管理义务。近年来,我国智能网联车辆、无人机等高风险人工智能系统应用不时引起危害后果,有必要推进人工智能安全管理义务相关立法,依法追究严重违反安全管理义务行为的法律责任,包括刑事责任。

三、人工智能安全管理义务承担者的刑事责任

目前我国人工智能安全立法适用范围窄,限于互联网信息服务领域的人工智能应用,法律位阶低,强制力弱,难以防控严重的和其他领域的人工智能安全风险。《互联网

^④ 根据《互联网信息服务深度合成管理规定》第16条和第17条的规定,互联网信息服务提供者应对其服务生成或编辑的信息内容进行标识,并向公众提示深度合成情况。根据《生成式人工智能服务管理暂行办法》第10条和第12条的规定,生成式人工智能服务提供者应对图片、视频等深度合成内容进行标识以及指导使用者。《人工智能生成合成内容标识办法》规定了具体的人工智能生成合成内容标识规范,提供网络信息内容传播服务的服务提供者、互联网应用程序分发平台应协助监督管理标识活动,使用网络信息内容传播服务发布生成合成内容的用户应承担声明和标识义务。

^⑤ 这里的系统性风险不同于人工智能系统的风险,特指通用人工智能模型因其高影响能力而具有的风险。参见欧盟《人工智能法》第3条第64、65款。

信息服务深度合成管理规定》等规章规定,违反其规定“构成犯罪的,依法追究刑事责任”,而现行刑法没有与违反人工智能安全管理义务犯罪相关的立法。人工智能安全是智慧社会发展的基础,应当给予充分的法律保障,有必要研究严重违反人工智能安全管理义务行为的刑事责任问题。

(一) 违反人工智能安全管理义务犯罪立法的正当性根据

违反人工智能安全管理义务犯罪立法具有必要性。人工智能引领新一轮科技革命和产业变革,对国家安全、经济社会发展和公众的福祉至关重要。《全球人工智能治理倡议》提出,发展人工智能应“以保障社会安全、尊重人类权益为前提,确保人工智能始终朝着有利于人类文明进步的方向发展”,要求发展“可审核、可监督、可追溯、可信赖的人工智能技术”。《中共中央关于进一步全面深化改革、推进中国式现代化的决定》要求,“必须全面贯彻总体国家安全观,完善维护国家安全体制机制,实现高质量发展和高水平安全良性互动”。人工智能安全是总体国家安全的重要组成部分,保障人工智能安全既不能“唯安全”,也不能“唯发展”,应以高水平安全来保障人工智能高质量发展。刑法应当贯彻前述政策,包容人工智能高质量发展,对安全风险采取底线控制,只要前置法能够防范安全风险失控,刑法就应坚守谦抑立场,坚持以行政监管为主、刑法威慑为辅。但是,如果人工智能的研发、生产、提供或使用危害国家安全、公共利益和公众合法权益,妨害人工智能安全管理秩序,那么其不仅不受刑法保护,还可能受到刑法的惩罚。

违反人工智能安全管理义务犯罪立法具有合理性。德国学者 Jakobs 教授将正犯分为支配犯和义务犯,支配犯因侵犯法益而具有可罚性,义务犯是违反为了建设良好的世界而设定的“制度管辖之义务”而成立犯罪。^④人工智能安全是智慧社会的基础,国家必然要建立人工智能安全管理制度,包括人工智能安全管理义务。为了保障以上制度的实施,遏制严重破坏人工智能安全管理秩序的行为,有必要设立违反人工智能安全管理义务犯罪,该类犯罪是违反人工智能安全管理制度所设定义务的义务犯。

违反人工智能安全管理义务犯罪是特殊的义务犯。德国学者 Roxin 教授认为义务犯具有对法益的“保护性控制”能力,^⑤ Schünemann、西田典之等教授认为,义务犯居于保证人地位,甚至将排他性支配作为犯罪成立的条件。^⑥例如,我国刑法中的拒不履行信息网络安全管理义务罪是义务犯,网络服务提供者创造网络空间,制定网络活动规则,有能力控制网络空间行为,居于排他性支配的保证人地位。违反人工智能安全管理义务犯罪与之不同,其义务承担者不具有此类排他性支配能力。由于人工智能的自主性及其风险控制的特殊性,人工智能安全相关人员仅具有不完全的保护性影响能力:(1) 由于人工智能的自主性和不可预测性,人工智能算法的形成和演变不完全受控于其研发者、提供者、使用者及最终用户,人工智能系统的自主性越高,就越不受控制;(2) 人工智能安全风险的形成和扩散分散于生产、流通、应用等多个阶段,各阶段上的义务承担者只能对

^④ 参见何庆仁:《义务犯研究》,中国人民大学出版社 2010 年版,第 29 页。

^⑤ 参见〔德〕克劳斯·罗克辛:《德国刑法学总论》(第 2 卷),王世洲等译,法律出版社 2013 年版,第 80、82 页。

^⑥ 参见前注^④,何庆仁书,第 65-68 页。

本阶段的安全风险发挥一定的防控作用,不能对风险进行排他性支配控制,其严格履行义务只能阶段性地、部分地、有限地贡献于人工智能安全。同样,其违反义务行为也只能阶段性地、部分地、有限地损害人工智能安全。按照风险分配原理和比例原则,只应承担有限的责任。因此,违反人工智能安全管理义务罪应有别于其他义务犯,有必要设置较高的入罪门槛和较轻的刑事责任,只将社会危害性严重的违反义务行为犯罪化,使该罪立法满足正当性要求。

(二) 违反人工智能安全管理义务罪的罪行构造

设立违反人工智能安全管理义务罪应当反映人工智能的特性,充分保障公民自由和合法权利,保障人工智能发展,有效防控人工智能安全风险。该罪是行政犯、义务犯和情节犯,应当审慎设置该罪的构成要件,除了必须具有严重的犯罪情节外,关键是合理规定人工智能安全管理刑法义务、主体和罪过要件。

1. 人工智能安全管理刑法义务的范围

该罪是义务犯,合理规定人工智能安全管理刑法义务对该罪具有重要意义。德国学者 Roxin 教授认为,义务犯是违反了构成要件之前的、刑法之外的特别义务的行为,公法、行业法规乃至民法上的义务都可以成为义务犯的特别义务。批评者认为,该观点是不合理的形式法律义务理论,^{④7}因为其他法律不以保护法益为目的,刑法义务只应从自己的任务中产生。后来 Roxin 教授将义务犯的特别义务限定为“与结果有关联地违反构成要件上的特别的义务”或“违反决定正犯性的义务”。^{④8}但是,如果将特别义务限定为与结果有关联,则必然将义务犯的范围缩小为结果犯。另外,所谓“决定正犯性”义务的限定则仍然没有明确特别义务的范围,反而有循环论证的逻辑谬误——义务犯以确定特别义务为前提,而特别义务的确定以成立义务犯为条件。德国学者 Jakobs 教授认为,义务犯的积极义务是团结义务,制度是强化团结的基础,违反团结义务决定义务犯的正犯性和可罚性。团结义务的范围并不从属于公法、民法等刑法之外的部门法规定的义务,而是与后者一样都来自于整体法秩序,刑法和其他部门法虽然有各自不同的任务,但都服务于维护整体法秩序。但是,刑法自身任务的独立性和独特性决定了义务犯的特别义务或积极义务不同于刑法之外的其他部门法规定的义务。Roxin 教授认为义务犯要回到法益侵害来确定正犯性,而 Jakobs 则认为犯罪的实质内容本来就是规范违反,导致二者对义务犯及其义务范围有不同的主张。^{④9}在德国刑法和我国刑法中,义务犯并非限于直接侵犯个人权益的犯罪,还包括侵害秩序法益的义务犯,如玩忽职守罪等。如果将集体法益排除在义务犯侵害的法益之外,则难以解释后一种义务犯。而按照 Jakobs 教授的规范违反论则难以区分其他部门法规定的义务。笔者认为,为了解决以上问题,应当回到刑法的任务,从义务犯设立目的出发讨论义务犯的义务范围。

^{④7} 参见许乃曼:《德国不作为犯学说的现状》,陈志辉译,载陈兴良主编:《刑事法评论》(第13卷),中国政法大学出版社2003年版,第390、391页。

^{④8} 参见前注^{④4},何庆仁书,第13、22页。

^{④9} 参见前注^{④4},何庆仁书,第30-31页。

设立违反人工智能安全管理义务罪的目的是保障人工智能安全发展,保护国家安全、公共安全和公众合法权益,应以此为出发点来确定人工智能安全管理刑法义务的范围。我国正在推进的人工智能立法和其他法律法规必然会规定安全管理义务,其中部分义务能成为刑法上的义务,违反此类义务的行为应当具有刑事可罚性。关于义务犯的刑事可罚性根据,德国刑法理论界和司法界主要采取风险降低理论,即“未实施行为的实施本来会(显著地)阻止结果发生的风险”,如果这种“降低风险的可能性”不仅事前能够判断,事后也可以确定,义务违反行为就具有刑事可罚性。^{⑤⑩}国内外人工智能安全立法规定了多种人工智能安全管理义务,例如,欧盟《人工智能法》规定了三类人工智能安全管理义务,即防控人工智能技术风险的义务、保障自然人基本权利相关的义务、监管程序要求的义务。^⑪我国前述规章也规定了多种安全管理义务。违反这些义务并非都具有刑事可罚性,判断以上各类义务违反行为是否具有刑事可罚性,评价标准应当是义务履行行为能否实质性、显著降低风险。遵守监管程序要求的义务并不具有直接降低风险的功能,违反该类义务不具有刑事可罚性,无论情节是否严重,都不应追究刑事责任。违反保障自然人基本权利相关义务的行为,能降低自然人权益遭受损害的风险,具有刑事可罚性。但是,其已受侵犯公民个人信息犯罪或其他侵犯公民人身权利、民主权利犯罪立法的独立规制,没有必要将其规定为违反人工智能安全管理义务犯罪。防控人工智能技术风险义务的履行能够实质性降低人工智能安全风险,如果在具体的场景下该类义务的违反与危害之间能建立起合法的连接,不仅能事前判断也能事后确定,该义务违反就具有刑事可罚性。

能够实质性降低严重人工智能安全风险义务可以成为人工智能安全管理刑法义务。欧盟《人工智能法》将人工智能安全风险类别化,针对不同类型和不同程度的人工智能安全风险设置不同的义务,值得我国相关立法借鉴。不过,该法规定的部分义务也有不合理之处。例如,该法针对非高风险的特定人工智能系统和通用人工智能模型,只规定透明度义务和其他更低约束力的义务,风险防控功效存疑。以生成式人工智能系统的安全管理义务为例,该类系统的提供者、部署者履行了标识和提示义务后,公众并非不受系统输出的虚假有害信息影响,仍然会有大量用户因为缺乏知识和经验以及对人工智能技术和大企业的盲目信任,受到系统提供的虚假有害信息欺骗而遭受严重损害,如佛罗里达14岁少年受谷歌聊天机器人诱导后自杀。^⑫在当前网络环境下,生成式人工智能系统的训练数据供给并非机会均等,对某个国家、民族、群体的数据歧视甚至敌视客观存在。^⑬如果对生成式人工智能系统只规定透明度义务,将无法有效防控其提供的违法信

^{⑤⑩} 参见前注^④,克劳斯·罗克辛书,第483-487页。

^⑪ 参见前注^②,皮勇文,第78页。

^⑫ 参见冬禾:《第一例AI机器人致死案,他已殉情》,载澎湃网, https://www.thepaper.cn/newsDetail_forward_29319515, 2026年2月6日访问。

^⑬ 参见郑海阳:《关于警惕ChatGPT滋生意识形态风险的建议》,载上海民建网, <https://www.mjshsw.org.cn/detailpage/jyxc-f6187d4e-cda5-4265-b14f-710d875d07a8.html>, 2026年2月6日访问。

息给我国国家安全、公共安全和秩序以及公众合法权益造成严重危害的风险。为了防控以上严重安全风险,我国对生成式人工智能系统的应用设置了风险防控义务。《生成式人工智能服务管理暂行办法》要求,提供者发现系统生成、传输违法内容的,“应当及时采取停止生成、停止传输、消除等处置措施,采取模型优化训练等措施进行整改,并向有关主管部门报告”。该义务履行是控制生成式人工智能系统生成、传播违法信息的风险实现和扩散的最后屏障,有必要将其确定为安全管理刑法义务,以刑法手段督促义务承担者切实履行。

基于以上分析,笔者建议,我国应通过制定人工智能相关法律法规,构建系统化的人工智能安全管理义务体系,将其中能够实质性降低严重安全风险的义务规定为人工智能安全管理刑法义务。该类义务的范围为:(1)高风险人工智能系统、信息服务类、情绪识别和生物识别分类人工智能系统的提供者,应当承担监控与降低安全风险、数据治理、算法治理、保持透明度、确保系统适合人工监督、确保系统的准确性和稳健性及网络安全的义务;(2)以上系统的部署者应当承担保持透明度、遵守使用说明、安排适合人员监督、监控和降低风险的义务。

2. 主体范围

违反人工智能安全管理义务罪的犯罪主体是人工智能安全管理义务的承担者,但是,该类义务的承担者并非都是该罪的主体。根据欧盟《人工智能法》的规定,高风险人工智能系统、非高风险的特定人工智能系统的提供者和部署者以及通用人工智能模型的提供者是主要的义务主体,出于风险防控的需要,高风险人工智能系统进口者、经销者以及欧盟境外提供者的授权代表被规定为协助监督、信息提供及与监管机构合作等义务的主体。我国《互联网信息服务深度合成管理规定》和《生成式人工智能服务管理暂行办法》将服务提供者规定为主要的义务主体,《互联网信息服务深度合成管理规定》还要求技术支持者承担数据管理、个人信息告知同意、算法安全审核与评估等义务(第14、15条),以上义务主体的范围比欧盟《人工智能法》更狭窄。以上主体并非都能成为违反人工智能安全管理义务罪的主体,成立该罪主体应当满足两个条件:一是承担人工智能安全管理刑法义务;二是其刑法义务的履行在客观上能够实质性、显著降低安全风险或者阻止风险实现,且因其义务违反导致风险的升高或实现,严重威胁人工智能安全及其他关联法益。否则,其义务违反行为就不满足刑事可罚性,只承担行政法律责任。

基于前述分析,笔者认为,违反人工智能安全管理义务罪的犯罪主体应当限于高风险人工智能系统及其他能够造成严重危害的人工智能系统的提供者和部署者。应注意,我国前述规章规定的服务提供者与欧盟《人工智能法》规定的人工智能系统或模型提供者不同,兼有后者规定的部署者地位,这里按后者规定的提供者和部署者来分析。人工智能系统的提供者是安全风险的引起者,主要承担人工智能研发阶段的风险控制义务,是从源头控制风险。而部署者承担应用阶段的风险控制义务,该阶段是风险实现阶段,也是阻止危害后果发生的关键环节,其义务履行对阻止风险实现的作用最大。二者都是主要的义务承担者,具有实际的风险防控能力,其义务履行对降低安全风险起关键

作用,可以成立该罪的犯罪主体。

人工智能的快速发展应用对国家竞争力、国家安全和经济社会发展具有重要意义,包容审慎成为人工智能法治理念,^{⑤4}人工智能刑事法治的首要任务应当是保障先进人工智能系统的研发和防范人工智能安全风险失控。为实现这个目标,对待以上两类主体要有所区别。部署者具有直接管控风险的能力(除非提供者未能提供有效的人工监管措施),在降低人工智能安全风险上具有“近因”作用,^{⑤5}对其义务违反导致的风险升高或失控具有预见能力,相比于提供者更符合归责条件。因此,对于部署者违反刑法义务的行为,应更多考虑保障安全的需要,合理认定其刑事可罚性。而提供者对促进先进人工智能发展的推动作用更大,应当更多地考虑促进先进人工智能发展的需要,保持包容的立场。当然,如果提供者兼有部署者身份,应当单独评价其作为部署者实施的行为。

人工智能系统的进口者、经销者、生产者、提供者的授权代表,以及通用人工智能模型的提供者和部署者不直接引起、扩大和实现风险,也不具有实质性降低风险的能力,一般情况下不应作为该罪的犯罪主体。欧盟《人工智能法》对以上主体规定的主要是监管程序性或预防性义务,这些义务履行只起到有限、间接降低风险的作用。我国前述规章未对这类主体规定安全管理义务,如果未来我国人工智能安全立法对其规定安全管理义务,对违反该类义务的行为只应追究行政法律责任。

3. 罪过形式

义务犯的罪过形式可以是故意或过失。欧盟《人工智能法》处罚故意或过失违反该法规定义务的行为(第99条第7款第i项)。违反人工智能安全管理义务罪的罪过形式是否包括故意和过失,影响该罪的处罚范围。笔者认为,故意违反人工智能安全管理刑法义务,且满足其他构成要件的,可以按该罪追究刑事责任;如果发生严重危害后果,对该后果无论是出于过失还是故意,都可以成立该罪,因为该罪处罚的是义务违反行为,因违反义务而未能阻止严重后果,即表明其行为具有严重的社会危害性;如果行为人对危害后果持希望或放任态度的,如自动驾驶汽车提供者在紧急路况处置算法中设置歧视性算法,使车辆选择性撞向行人,部署者故意违反自动驾驶汽车人工监管义务,放任车辆撞死路人或乘客的,可以同时构成故意杀人罪、以危险方法危害公共安全罪等故意犯罪,应按想象竞合犯从一重罪定罪,从重处罚。

行为人应当知道会违反以上义务,但因疏忽大意而不知道有此义务且未履行,如果将该种心态作为该罪的罪过形态,将扩大该罪的处罚范围,其合理性值得讨论。人工智能安全是独立的新公共法益,同时关联着国家安全等重大法益,已成为总体国家安全的新组成部分并深度嵌入公共安全,应属于重大法益,为其提供刑法保护具有合理性。既然《刑法》处罚过失违反国家秘密、军事秘密保密义务的行为,也可以将违反人工智能安全管理义务犯罪扩展到过失犯罪。但是,从保障人工智能发展的角度看,这样对人工

^{⑤4} 参见黄籍:《论我国人工智能领域包容审慎监管的法治维度》,载《财经法学》2025年第2期,第98-100页。

^{⑤5} 参见[美]约书亚·德雷斯勒:《美国刑法精解》(第4版),王秀梅等译,北京大学出版社2009年版,第171-173页。

智能相关方处罚过重,可能阻碍人工智能的研发、生产和应用,不符合促进人工智能快速发展应用的国家战略。

基于对人工智能安全与发展的考虑,笔者认为,该罪应当限于故意犯,可以通过协调人工智能安全相关行政、刑事立法和执法,消除对过失违反义务行为追究刑事责任的可能性。可以通过加强人工智能安全监管执法,督促义务承担者履行安全管理义务,减少或排除行为人过失违反义务的可能性,因而没有必要将极少发生的过失违反义务行为设定为犯罪。如果行为人曾经因过失违反人工智能安全管理刑法义务受到行政处罚,可以推定行为人知悉应当履行的义务,其再次违反义务就只可能是出于故意,此时刑法所处罚的是故意违反义务行为,而非过失违反行为。在罪状设计上,可以参考拒不履行信息网络安全管理义务罪,将该罪的行为规定为“拒不履行安全管理义务,情节严重的行为”,并通过增加监管机关责令整改的程序,排除行为人过失违反义务的情形。

Abstract: In cases where autonomous decision-making and control by AI systems lead to harmful consequences, there are gaps in the current laws. Civil and administrative legal measures are insufficient to curb such severe risks, necessitating criminal law governance. Regarding the approach of criminal law governance, legislation on crimes that disrupt the order of AI safety management can balance both AI safety and development. Among these, adopting a legislative plan for the crime that violates AI safety management obligations can curb the harms. AI safety management obligations are specialized and limited regulatory responsibilities designed to prevent and control AI-related risks. To effectively penalize serious violations of the obligations, it is necessary to establish a crime that violates AI safety management obligations. Based on criminal law theories such as duty-based crime and risk allocation, this crime should be reasonably defined in terms of AI safety management obligations, subjects, and culpability to effectively mitigate AI-related risks.

(责任编辑:王 楷)