中外法学 Peking University Law Journal
Vol. 37, No. 2(2025)pp. 285－305

高风险人工智能的法律界定及规制

解志勇*

摘 要 人工智能技术的底层科学逻辑，决定了其风险可认知、可描述、可分析、可界定。与欧盟《人工智能法》确立的一般风险治理进路不同，我国可采取“风险＋情境融合治理”的进路，以高质效应对高风险人工智能的系统风险。“高风险”的界定标准，可基于人工智能本体能力强弱、功能作用对象、潜在致害程度等三个维度进行确认。首先，风险级别与人工智能本体强弱呈正相关关系，强人工智能与超强人工智能均属高风险系统；其次，直接作用对象为“重大安全”，涵盖国家安全、社会公共安全、个人生命安全以及其他重要基本权利安全；再次，存在对“重大安全”造成实质性显著减损的可能。鉴于此，高风险人工智能的治理与规制，主要是在特定情境中针对其安全性展开，应遵循合法性原则、科技伦理约束原则、技术治理原则，秉持包容审慎与合作规制理念。以情境作为治理单元，以安全维护为首要目标，采取“行为规制＋个体赋权”立法模式，进行事前、事中、事后的全过程治理，立法上最终达致制定《人工智能安全法》的目标。

关 键 词 人工智能 高风险人工智能 人工智能安全法 治理 规制

一、引言：人工智能普遍应用的价值和风险

习近平总书记指出，“人工智能是新一轮科技革命和产业变革的重要驱动力量，将对全球经济社会发展和人类文明进步产生深远影响”。〔1〕鉴于人工智能普遍应用带来的促进经济效益、社会进步、文明更迭等价值，拥抱人工智能已成为时代所趋。但人工智能的普遍应用也

* 中国政法大学比较法学研究院教授。本文系国家社科基金重大项目“应对新一轮科技革命的法治体系完善与基本法理研究”(项目编号：24&ZD134)的阶段性研究成果。

〔1〕《习近平向2024世界智能产业博览会致贺信》，载中国政府网，https://www.gov.cn/yaowen/liebiao/202406/content_6958352.htm，最后访问日期：2024年12月3日。

会给国家安全、社会安全、个人安全等带来挑战，亟须立法作出有效回应。

2024 年 3 月 13 日，全球首部人工智能监管法规——欧盟《人工智能法》(Artificial Intelligence Act)的通过，掀起了各国对人工智能规制路径与模式选择的新一轮讨论热潮。该法案以防范人工智能风险为主要定位，将高风险人工智能系统作为主要规制对象。法案首先以损害发生可能性和损害严重程度的结合作为风险定义，对“高风险”人工智能系统用途采用“抽象要件＋明列领域”的方法加以界定；其次，根据“对自然人的健康、安全或基本权利构成重大损害风险”的叠加限制条件，具体认定是否构成“高风险”用途场景；再次，对构成高风险用途的，匹配以更为严苛的规制规范。我国在建构人工智能风险治理体系时，可借鉴其合理成分，但不可照搬照抄。本文尝试建构“高风险”界定标准，这种“高风险”，若基于风险的规制模式，接近于“不可接受的风险”，也即因其风险程度很高或损害结果难以令人接受，所以必须采取预防性的干预措施。但这种“不可接受的风险”，若从欧盟《人工智能法》切入，并不属于禁止的风险，在发生情境中又更多与法案中的“高风险”重合，故本文倾向于使用“高风险”概念，来概括归纳这种产生风险属于“不可接受”但又不能全然禁止的人工智能系统，并在此基础上提出高风险人工智能系统的具体治理进路和规制方案。

二、人工智能风险的一般治理进路

对人工智能采取简单化风险治理进路，可能存在不够全面周延的问题，但若能采取“风险＋情境融合治理”的进路，便可大大提高风险治理的质效。

(一)欧盟的风险治理进路

欧盟《人工智能法》所采取的风险治理逻辑有三步：一是将风险界定为“损害”；二是对可能引发风险的行为进行评估；三是划分风险类别，并配置相应的责任规范。此种风险治理进路，可能面临多种质疑。

1. 风险评估困难

有观点指出，欧盟采取的人工智能风险分级治理的科学性与可行性存疑。〔2〕风险管理以有效的风险评估为前提，而风险评估常用的就是成本效益分析方法。这种传统方法一方面在欧盟《人工智能法》施行之际并未得到充分运用，具有代表性的批评意见是其风险治理进路只片面考虑了风险，未考虑收益，威胁创新；〔3〕另一方面也与人工智能风险或不确定性的“不可度量性”相冲突。〔4〕同时，对人工智能活动引发的风险进行量化分析殊为不易。风险的量

〔2〕 参见丁晓东：“人工智能风险的法律规制——以欧盟《人工智能法》为例”，《法律科学》2024 年第 5 期，第 10—12 页。

〔3〕 See Karen L. Neuman, Hilary Bonaccorsi, Michael P. Tierney and Madeleine White, “European Commission's Proposed Regulation on Artificial Intelligence: Requirements for High-Risk AI Systems,” *The Journal of Robotics, Artificial Intelligence & Law*, Vol. 4, No. 6, 2021, pp. 441-449.

〔4〕 参见(美)弗兰克·H. 奈特：《风险、不确定性与利润》，安佳译，商务印书馆 2006 年版，第 211 页。

化分析，通常需要掌握损害的危害程度、发生概率、分布普遍性、持续性、可逆性等基本数据，〔5〕在欠缺高质量数据、与新技术匹配的有效量化模型时，精确计算出最终的风险量化结果是很困难的。〔6〕即使在不考虑量化的前提下，风险性质的评估把握也是一大难题，比如当涉及通用人工智能模型时，其构成“系统性风险”的界定标准也难以捉摸。

2. 风险分类遗漏

普通的风险治理进路，采取风险的静态分级，存在较为机械和分类遗漏问题。首先，就产品构成类的人工智能系统而言，产品的风险高低与作为其构成部分的人工智能的风险高低，并无必然联系，分级缺乏直接指导意义。其次，就独立的人工智能系统而言，欧盟《人工智能法》附件三所列举的用途范围，也存在未将其他属于高风险的人工智能系统列入规制范围的情形，比如侵入性极强的诊疗人工智能系统。再者，即使根据《人工智能法》第 7 条，欧盟委员会有权增加或修改高风险人工智能系统用例，也有权在附件三中的人工智能系统不再对基本权利、健康与安全造成显著风险时，将其排除在规制范围外。上述举措仍无法从根本上确保人工智能分级的全面周延，因为大量不确定概念的堆砌、叠加，带来了更大的弹性解释空间，最终可能演变为“政治决策的包装”。〔7〕

3. 实施成本高昂

前述风险治理进路，还伴随着高昂的实施成本。对于人工智能这一类新兴技术，选择恰当的规制时间点较为不易。过早规制可能产生抑制科技创新发展的后果，但若等到人工智能技术充分发展以后再行介入，此时的技术模式已经成熟、固定，对其进行校正将会面临不小的成本，也可能因研发、提供、使用者等多方反对而难以推行。〔8〕而从规制主体所需具备的能力来看，因规制者相较科技企业，明显处于信息与技术劣势，若想要实现有效的全过程监管，提升分析、评估、控制人工智能风险的能力，也需耗费巨大的人力、物力、财力资源。

4. 立法话语差异

我国与欧盟的立法话语存在差异。欧盟人工智能立法的发展思路，是从统筹规划欧盟统一大市场、弥合各国分歧的大背景出发，然后再确立具体制度，这与其数字战略、建立统一数字市场的目标息息相关。而我国在人工智能领域采取的是从具体到综合的立法模式，与其方法策略迥异。同时，中国在数字时代对技术的管理模式，一定程度上带有“国家驱动”色彩，〔9〕

〔5〕 See Ortwin Renn and Andreas Klinke, Risk Governance, “Concept and Application to Technological Risk,” in Adam Burgess, Alberto Alemanno and Jens Zinn (eds.), *Routledge Handbook of Risk Studies*, London: Routledge Press, 2016, pp. 204-215.

〔6〕 参见林洹民：“论人工智能立法的基本路径”，《中国法学》2024 年第 5 期，第 85 页。

〔7〕 参见王天凡：“人工智能监管的路径选择——欧盟《人工智能法》的范式、争议及影响”，《欧洲研究》2024 年第 3 期，第 11 页。

〔8〕 See Michael Guihot, Anne F. Matthew and Nicolas P. Suzor, “Nudging Robots: Innovative Solutions to Regulate Artificial Intelligence,” *Vanderbilt Journal of Entertainment & Technology Law*, Vol. 20, No. 2, 2017, pp. 385-456.

〔9〕 See Anu Bradford, *Digital Empires: The Global Battle to Regulate Technology*, Oxford: Oxford University Press, 2023, p. 80.

这表明，相较于欧盟，在我国的本土制度环境中，立法更易发展为国家治理意愿驱动下的制度样态。而从事人工智能研发、提供、使用活动的科研机构与企业，在与政府的互动关系中，也更易受到政府影响并采取相应的行动机制。

(二)人工智能风险的可界定性

人工智能是计算机科学的一个重要分支，是模拟、延伸和拓展人类智能活动的科学，〔10〕基于此底层科学逻辑，其风险也应当是可认知、可描述、可分析的，具有可界定性，有必要继续坚持将风险分级作为界定人工智能系统风险的基础工具。

1.人工智能风险的可认知性

人工智能风险本身具有的不确定性、不透明性、变动性和复杂性等特点，并不影响其仍具有可认知性。比如当前对人工智能风险的认知分类，就有观点将其分为技术风险、社会风险、伦理风险等技术风险如“AI 数据隐私与安全”“算法误导”“算法歧视”“算法黑箱”“算法滥用”“算法偏差”等；有关社会风险的认知如“信息茧房”“替代风险”等；伦理风险有“弱价值风险”“去伦理风险”等。〔11〕虽然以上分类仍处于一个初步展开的状态，但不可否认随着可解释 AI 和透明算法的研究，越来越多的人工智能风险可以被人类认知、发现和理解。

2.人工智能风险的可描述性

在人工智能风险可认知的基础上，人工智能风险还具有可描述性且日益精确。人工智能系统的风险，迫使立法者基于功利主义、后果论对其进行事前规制，但风险的事前分级并不排斥具体个案风险评估对于情境因素的描述，比如通过事前建立动态调整的高风险人工智能系统的情境目录，以形成更加精细化的风险描述图景。而情境目录的后续更新，来源于不断积累的经验法则和技术标准，两者能够弥补技术发展过程中可能产生的原初分类描述疏漏问题。一方面，经验中所包含的各种可能性是规范所描述的风险的主要来源；〔12〕另一方面，特定风险技术标准的形成与完善，能够对风险描述指标进行评分乃至加权运算，〔13〕为描绘后续技术发展过程中出现的新风险筑牢基石。

3.人工智能风险的可分析性

成本效益分析方法除了上文提及的可能面临风险无法评估的问题，还可能忽略个体对风险可接受程度的差异。〔14〕但这种方法的“失灵”，并不代表人工智能风险不可分析。风险分析是一个渐进和动态的过程，随着技术进步和数据积累，风险分析模型可以逐步完善，比如利

〔10〕 参见周辉、徐玖玖、朱悦、杨心宇：《人工智能治理：场景、原则与规则》，中国社会科学出版社 2021 年版，第 3 页。

〔11〕 王海明：《法治化防控人工智能风险》，载中国社会科学网，https://www.cssn.cn/skgz/bwyc/202411/t20241115_5802600.shtml，最后访问日期：2024 年 12 月 8 日。

〔12〕 参见(美)杜威：《杜威五大讲演》，胡适译，安徽教育出版社 2005 年版，第 236 页。

〔13〕 参见(美)道格拉斯·W. 哈伯德：《风险管理的失败：为什么失败以及如何修复》，杨雪蕾、蔡松儒、王丽洁译，中国金融出版社 2021 年版，第 25 页。

〔14〕 参见(美)玛丽·道格拉斯：《风险的接受：社会科学的视角》，熊畅译，华东师范大学出版社 2022 年版，第 47—51 页。

用奥地利学者威尔伯格(Walter Wilburg)提出的动态系统论,[15]可以重新建构一套由多个弹性要素构成的人工智能风险预估动态构造。在此认识基础上,尽管有些风险不易量化、把握性质,但仍可以采用多维度的评估方法来对其进行级别定性分析。例如,通过级别界定维度的建构,将高风险人工智能系统风险分解为能力维度、功能维度、受损维度,并依次分析这些维度上的风险。是故,为风险分级评估保留一定的拓展空间,仍有必要。

(三)"风险+情境融合治理"进路

在坚持风险分级的基础上,以回应规制理论、情境化方法论为指导,我国可以采取"风险+情境融合治理"的进路,以应对高风险人工智能系统的潜在风险。

1.风险社会理论的要求

德国社会学家乌尔里希·贝克将"风险"的概念深入到社会学研究领域,并结合自反性现代化、系统脆弱性、全球性等理论,演绎生成著名的风险社会理论,该理论被视为分析风险、不确定性及其社会化过程的典型理论范式。[16] 而高风险人工智能系统的风险,实质上就是指数据异化、算法黑箱、算力垄断可能带来更高层级的现代安全风险、自毁性安全风险、分配性安全风险等,其安全风险服从风险社会理论的分配逻辑,也可与风险社会理论高度耦合,这也正是将此传统理论引入高风险人工智能系统安全风险防范化解研究的原因所在。[17]

2.回应规制理论的进化

风险规制模式虽有多种类型,但从对待风险的态度以及所采取的规制策略上来看,整体上可分为预防原则模式和基于风险的规制模式两类,后者又进一步衍生出真正的回应式基于风险的规制模式。[18] 回应型规制设想规制活动发生于交互对话的环境中,规制者通常优先采用干预性较低的措施,当这些措施失灵则逐步采取更具有惩罚性或强制性的措施。这类规制最突出的特点即后果的动态回应性,[19]旨在弥合强监管和放松监管之间的鸿沟。相较低风险人工智能系统而言,高风险人工智能系统需要更为精细化地确定事前规制的目标、确定风险容忍度、进行风险评估,并根据风险评估分配规制资源,从而提高规制效率。故有必要对其采纳一种及时、回应性强的积极规制观,进行动态适应和有力引导。

3.情境化方法论的指导

对人工智能风险所涉及的领域乃至威胁的权利类型的识别与控制,不可避免地会涉及具体情境中的利益权衡和价值判断,以致一般的风险治理进路无法完全解决——同一项技术在

---

〔15〕 该理论的基本构想是,法律后果的形成,是基于多种动态作用力的"协同作用",参见(奥)瓦尔特·维尔伯格:"私法领域内动态体系的发展",李昊译,《苏州大学学报(法学版)》2015年第4期,第112—114页。

〔16〕 参见(德)乌尔里希·贝克:《风险社会》,何博闻译,译林出版社2004年版,第7页。

〔17〕 参见陈嘉鑫、李宝诚:"风险社会理论视域下生成式人工智能安全风险检视与应对",《情报杂志》2025年第1期,第128—135页。

〔18〕 See Julia Blackand Robert Baldwin, "Really Responsive Risk-Based Regulation," *Law & Policy*, Vol. 32, No. 2, 2010, pp. 181-213.

〔19〕 See Ian Ayres and John Braithwaite, *Responsive Regulation: Transcending the Deregulation Debate*, New York: Oxford University Press, 1992, pp. 35-37.

不同行业、领域应用产生的性质、程度不同的风险，也即人工智能系统嵌入性特征带来的规制困境。〔20〕就此，学界不少学者提出我国应采取基于“场景”的人工智能风险规制模式。本文在此更愿意将其称为基于“情境”的人工智能风险规制模式，因为“情境”的内涵相较“场景”更为广泛，且虽“场景”有明显的技术塑造特征，但正如上文所述，实质上人们所接受、可容忍的风险，至少在决策者看来应该是最佳的选项，〔21〕而不是技术风险本身。故“情境化方法论”才是决策者应对人工智能风险规制中“无知”或者“不确定性”的最有效方法。〔22〕

三、“高风险”人工智能系统的界定标准及主要情境

在坚持风险分级规制进路的前提下，如何破解“概率＋严重程度”的组合指标在高风险人工智能系统运行时难以计算的困境？答案是可以将“高风险”的界定标准，纳入人工智能本体能力强弱、功能作用对象、潜在致害程度三个维度，再将其置于主要情境中进行风险具体分析。

（一）能力维度：强弱本体

界定标准中的首要维度为人工智能本体的能力强弱。人工智能按照其本体强弱的能力区分，可分为弱人工智能、强人工智能和超级人工智能三类。其中，弱人工智能专注于完成特定任务，仅针对提高生产力与经济效益的某一特定问题域进行了辅助优化。强人工智能是拥有广泛认知能力的机器，能够理解和学习几乎任何人类智能所能做的事情，“与人类心灵等价”。〔23〕而超级人工智能作为强人工智能的更极端版本，目前只在理论上存在，其智能水平远超最聪明的人类大脑，能够解决极其复杂的任务，并进行自我改进。〔24〕人工智能系统的风险高低程度，与其本体能力的强弱呈正相关的关系，一般而言，人工智能本体越强，产生的风险级别越高。这一论断背后的逻辑是，人工智能本体系统技术能力的增强，使其决策范围、自动化程度、任务复杂度、自我优化与提升能力等均大大提升；而在这些本体属性增强的同时，人工智能潜在风险的不可预测性、逸脱人类控制可能性、错误带来的风险放大性、风险规模也相应增加。鉴于此，强人工智能与超级人工智能，原则上均属高风险系统；而弱人工智能是否属于高风险系统，还需结合具体情境加以讨论。

（二）功能维度：重大安全

界定标准中的第二个维度为功能维度，意指高风险系统的直接作用对象为“重大安全”。

---

〔20〕 参见丁晓东：“全球比较下的我国人工智能立法”，《比较法研究》2024 年第 4 期，第 59 页。

〔21〕 参见（英）巴鲁克·费斯科霍夫、莎拉·利希藤斯坦、保罗·斯诺维克、斯蒂芬·德比、拉尔夫·基尼：《人类可接受风险》，王红漫译，北京大学出版社 2009 年版，第 3—11 页。

〔22〕 See Charles F. Sabel and William H. Simon, “Contextualizing Regimes: Institutionalization as a Response to the Limits of Interpretation and Policy Engineering,” *Michigan Law Review*, Vol. 110, No. 7, 2012, pp. 1265-1308.

〔23〕 参见（美）约翰·塞尔：《心、脑与科学》，杨音莱译，上海译文出版社 2006 年版，第 417—424 页。

〔24〕 参见（英）尼克·波斯特洛姆：《超级智能：路线图、危险性与应对策略》，张体伟、张玉青译，中信出版社 2015 年版，第 29—30 页。

“安全”的基本概念是指“免除了不可接受的损害风险的状态”,〔25〕是一系列特定的危险被消除或者被降低到最低限度的一种情景。〔26〕“重大安全”并不局限于公领域的公共安全,〔27〕而是涵盖国家安全、社会公共安全、个人生命安全以及其他重要基本权利安全。

1.国家安全

从实证法的角度出发,“国家安全”是指国家政权、主权、统一和领土完整、人民福祉、经济社会可持续发展和国家其他重大利益相对处于没有危险和不受内外威胁的状态,以及保障持续安全状态的能力。〔28〕若进一步细化,“国家安全”在宏观上可被划分为人民安全、政治安全、经济安全以及军事、文化安全等模块。〔29〕我们需要警惕高风险人工智能系统,尤其是生成式人工智能大模型在运转过程中,因应用产生偏见与歧视等有害内容以及泄露有关国家安全的敏感数据滥用于欺骗操纵等违法犯罪行为,〔30〕这会对国家安全尤其是政治安全、文化安全等造成严重损害。

2.社会公共安全(含群体安全)

“社会公共安全”的独立性,体现在社会安定、社会秩序良好、人民安居乐业的重要保障层面。时至今日,社会公共安全仍是一个不确定法律概念,从其利益表现形式——“公共利益”的内容、受益对象的模糊性中也可窥见。〔31〕举例而言,高风险人工智能系统在自动驾驶情境可能对交通驾驶秩序安全造成破坏,在新闻服务情境可能对社会舆论安全造成破坏。“社会公共安全”并不排除“群体安全”,因为“公”的存在本身就是离不开群体的,“社会公共安全”也可被视为一种连接个人与群体之间的纽带。〔32〕因此,当人工智能系统的运行,可能显著威胁到“数量上达多数”的共同体利益时,〔33〕也应被视为对社会公共安全的破坏,应纳入高风险规制范围之中。

3.个人生命安全

国家和社会公共安全之外的非公共安全,也有属于“高风险”人工智能系统可能造成显著减损的重大安全类型,最典型的就是个人生命安全。个人生命安全在权益层面表现为生命权,基于生命之神圣性,在人权中具有优位性。〔34〕作为最高的人格利益,其以自然人的生命安全

〔25〕 参见管理科学技术名词审定委员会编:《管理科学技术名词》,科学出版社2016年版,第462页。

〔26〕 参见(英)安东尼·吉登斯:《现代性的后果》,田禾译,黄平校,译林出版社2011年版,第31页。

〔27〕 参见欧盟《人工智能法》第1.1条,其将保护私主体的健康、基本权利作为规制人工智能风险的主要目的;还有如美国、英国、欧盟等多方共同签署的《人工智能与人权、民主、法治框架公约》所关注的人工智能风险,同样也包括人工智能系统在整个生命周期内对人权所造成的不利影响。

〔28〕 《中华人民共和国国家安全法》第2条。

〔29〕 《中华人民共和国国家安全法》第3条。

〔30〕 参见刘金瑞:“生成式人工智能大模型的新型风险与规制框架”,《行政法学研究》2024年第2期,第19—21页。

〔31〕 参见陈新民:《德国公法学基础理论(上册)》,山东人民出版社2001年版,第182页。

〔32〕 参见李明伍:“公共性的一般类型及其若干传统模型”,《社会学研究》1997年第4期,第109页。

〔33〕 参见胡锦光、王锴:“论我国宪法中‘公共利益’的界定”,《中国法学》2005年第1期,第19页。

〔34〕 参见解志勇:“生命伦理法的建构”,《比较法研究》2024年第1期,第10页。

利益为内容，是自然人的人格载体，是自然人享有的其他所有权利和利益的基础和前提。〔35〕

4. 其他重要基本权利安全

重大的非公共安全除了个人生命安全，还包括其他个人基本权利的安全，但仅限于健康、自由、平等、人格尊严、隐私等位阶较高的重要基本权利，而不包括纯粹的财产权等经济上权利。比如当人工智能系统运用于专家系统，像对信息准确性要求较高的诊疗行业，患者如果轻信人工智能系统就某些身体不适症状给出的错误治疗建议，不去就医或错误服用药物剂量等，就可能会延误救治而对身体健康权造成伤害；再如人工智能的偏见和歧视问题可能导致对特定群体的不公正对待，违反非歧视和公平原则，造成特定群体无法获得社会保障权利。

（三）受损维度：显著减损

界定标准的第三个维度为受损维度，高风险人工智能系统应当存在对“重大安全”造成实质性显著减损的可能。这种潜在致害程度，可从对重大安全等所造成减损的实质影响、风险规模大小等因素加以认识。〔36〕

1. 实质影响

并非所有未来风险均构成损害，只有那些有据可循的“实质性风险”（substantial risk）才具有相当程度的可实现性，高风险人工智能系统对安全的威胁认定也需要符合“实质影响”标准。〔37〕当然，这种威胁是否是实质性的，有赖于个案中综合考虑各种因素，〔38〕必须区分那些毫无根据的主观臆测与客观合理的风险。鉴于高风险人工智能系统所涉安全的层级较高，一旦致损可能带来毁灭性的后果，对此处“实质影响”的认定标准，宜作宽容解释。只要存在对“重大安全”造成显著减损的一定可能，就予以认定构成“实质影响”，不要求造成显著减损的高度盖然性，而这种存在可能性的结论应当由人工智能主管部门通过经验结合技术标准审慎作出。

2. 风险规模

高风险人工智能系统的认定，需要满足其对重大安全造成的“风险规模巨大”的标准。这一标准，从系统自身出发，需要关注人工智能基础模型的算力、参数、使用规模，当规模超过一定技术临界值，即构成巨大规模风险；从系统作用对象出发，需要关注被作用方的涉众性，还有被作用方对人工智能系统输出结果的依赖程度、交互对等关系、损害的可逆性等因素，若具有舆论属性、社会动员性，或者说潜在受害者属于极度依赖人工智能输出的结果，在两者交互关系中属于被不平等对待方，损害一旦发生便处于不可逆的境地时，也应认为构成巨大规模风险；此外，还可从系统运作全过程出发，结合其预期用途、已经使用的程度、持续时间等因素进行综合考量。

---

〔35〕 参见王利明：《人格权法研究》，中国人民大学出版社 2005 年版，第 303 页。

〔36〕 参见赵鹏：《风险社会的行政法回应》，中国政法大学出版社 2018 年版，第 10 页。

〔37〕 See Attias v. CareFirst, Inc., 865 F.3d 620, 627 (D.C. Cir. 2017).

〔38〕 See Jameson Steffel, “The Time between the Theft and the Injury: Standing Requirements Based on a Future Risk of Identity Theft after a Data Breach,” *University of Cincinnati Law Review*, Vol. 88, No. 4, 2020, p. 1189.

基于上述三个维度，当本体较强的人工智能系统存在对“重大安全”造成实质性显著减损的可能，这种潜在致害的不利益性状态就是人工智能系统分类中“高风险”的主要界定标准。当然，对上述界定标准还应设置一定的“白名单”加以辅助。人工智能系统若不会对重大安全构成实质性显著减损程度的破坏，而只是试图在限定范围内履行某项程序性任务，或旨在改善先前人类已完成的活动，或只是试图对已有决策模式加以检测，而并不打算取代或影响此前已完成的人类评估决定，或只是准备承担预备性任务，则均不应纳入“高风险”范畴进行规制。

(四)主要情境

在情境化方法论下，“高风险”人工智能系统的界定必须具有情境内嵌性的特征，即注重“重大安全”与情境的融合观察，在此主要列举九类情境。

1.关键数字基础设施

随着人工智能技术在社会经济生活中的广泛应用，特别是在不同行业关键基础设施中的应用，越来越多的领域可能产生重大风险，有可能导致大规模的、难以恢复的、社会难以接受的国家公共安全破坏。我国《关键信息基础设施安全保护条例》明确的规制范畴，主要聚焦公共通信和信息服务、能源、交通、水利、金融、公共服务、电子政务、国防科技工业等重要行业和领域的，以及其他一旦遭到破坏、丧失功能或者数据泄露，可能严重危害国家安全、国计民生、公共利益的重要网络设施、信息系统等。〔39〕

2.生物识别

生物识别是指对人类指纹、面部、步态等不同类型的个人身份数据进行收集、处理、分析，并以此为基础识别、跟踪特定对象，产生的“生物识别信息”可作为识别个人的“唯一标识”。〔40〕此类技术应用情境中，一旦技术失控或者被滥用，可能导致非公共安全中的其他重要基本权利，如个人信息受保护权这一基本权利受损。〔41〕

3.教育就业

考虑到劳动权和受教育权均同时具有自由权和社会权面向，〔42〕此处将其作合并处理，嵌入高风险予以讨论。涉及教育、就业的情境主要包括以下两类。一种是当人工智能系统用于确定自然人进入各级教育和职业培训机构或课程的机会、录取或分配，用于评估学习成果，以及在教育和职业培训机构内用于评估教育水平、监控监测违纪行为之时，可能会产生加剧教育、职业不平等风险以及隐私被侵犯的重大安全减损。另一种情形则是随着人工智能系统在人力资源管理中的广泛使用，员工的晋升、绩效管理和解雇等问题均可由系统预测并输出决策结果，但这种技术预测发生在员工难以理解算法背后逻辑的“黑箱”运转中。其中可能嵌入的歧视性偏见，会影响员工的工作机会和工资薪酬，影响个体平等就业权等自我发展权利的实

〔39〕 参见我国《关键信息基础设施安全保护条例》第2条。

〔40〕 参见付微明：“个人生物识别信息的法律保护模式与中国选择”，《华东政法大学学报》2019年第6期，第78页。

〔41〕 参见王锡锌：“国家保护视野中的个人信息权利束”，《中国社会科学》2021年第11期，第122页。

〔42〕 参见陈征：“宪法社会权的价值属性与规范定位”，《环球法律评论》2024年第5期，第7页。

现,甚至对劳动力市场上的平等秩序造成破坏。

4.自动驾驶

在现有自动驾驶尚未达成高度智能化决策水平的情况下,自动驾驶人工智能系统应用,主要涉及的风险是行车过程中遭遇复杂路况时,自动驾驶车辆能否基于尊重生命安全和人格尊严的考量作出最合理的行车决策。这种决策结果的输出,不仅可能威胁个人生命安全以及其他重要基本权利安全,也可能因对其他交通参与者的不合理差别对待导致社会公平受损,有必要纳入高风险规制范围。

5.诊疗服务

诊疗人工智能系统直接应用于诊疗活动,与患者的生命健康权紧密关联,具有更强的涉人身属性和侵入性。〔43〕此类系统主要依赖算法模型进行计算,患者无法理解基本原理和输出决策的可靠性、合理性;囿于医疗措施的复杂性,医务人员在很多情况下也难以充分尽到解释说明义务;而算法的研发者以及后台控制者,一般又缺乏医学方面的专业知识,容易受限于自身主观价值倾向。当前利用"知情—选择决定"三者一体的法律技术,来分配诊疗损害责任,其实是法律解释不了造成风险的真正来源而作出的无奈推理。对于此类情境,需要将其纳入高风险人工智能系统规制的范围,并寻求高风险规制规范与产品责任的协调。

6.新闻服务

在利用人工智能系统提供互联网新闻信息服务的情境中,鉴于此情境的涉众性极高,一旦丧失作为国家安全"晴雨表"的舆论安全,〔44〕将直接或间接催生出诸如意识形态对立、阶级对立等国家安全减损态势。此外,新闻服务中也可能因错误、非准确信息传播,抑或无关真假的意见表达而引发群体安全减损,比如借助社交媒体中的社交机器人言论,〔45〕在算法扩大逻辑与"社会流瀑效应"的双重加持下,就极有可能持续刺激群体产生恐慌不安乃至暴力倾向,严重破坏社会公共安全。

7.授益给付

当将人工智能系统引入应用于行政给付领域,用于确定行政机关是否应给予、拒绝、削减、撤销或收回相应的津贴与公共服务时,一旦系统输出否定性结果,可能会给当事人的生计乃至其他重要基本权利带来严重影响,此时该类系统也应被列入高风险之列。而随着给付行政理论的扩张,涉及公民生存利益、发展利益和共享利益的给付类人工智能系统,〔46〕都应被囊括在规制情境之中。与此同时,人工智能系统也在能否获得私营服务的判断中广泛应用,比如决定个人健康和人寿保险资格的系统、通过评估自然人信用分或信誉度而后决定其能够获得金融服务资格的系统等。鉴于这些系统输出的私营服务资格有无、优先顺序,都会对个人重要基

〔43〕 参见郑志峰:"诊疗人工智能的医疗损害责任",《中国法学》2023 年第 1 期,第 204 页。

〔44〕 参见许加彪、王军峰:"算法安全:伪舆论的隐形机制与风险治理",《现代传播(中国传媒大学学报)》2022 年第 8 期,第 138—146 页。

〔45〕 李晟:"国家安全视角下社交机器人的法律规制",《中外法学》2022 年第 2 期,第 431 页。

〔46〕 参见解志勇:"基于中国式扶贫实践的给付行政法治创新",《法学研究》2022 年第 6 期,第 23—24 页。

本权利造成实质影响，也有必要将其纳入高风险之列。

8.司法程序

除上述运用人工智能系统辅助执法的情境，使用人工智能系统辅助司法程序的情境，也当纳入高风险范畴。此处的辅助“司法程序”，既包括作用于可能影响自然人或组织成为刑事犯罪评价对象的程序，如用于预测刑事犯罪的可能性、调查欺诈性内容等证据评估的人工智能系统；同时，也包括作用于协助司法机关研究和解释事实和法律、将法律适用于某一组具体事实或以类似方式用于替代性争议解决的人工智能系统。上述情境于个人，可能导致辩方当事人的辩护权受损；于社会秩序，也可能因为系统刻板印象冲击公正价值，损害公正法治秩序。〔47〕

9.通用人工智能模型

区别于以上八种特定情境，通用人工智能模型并不用于解决特定情境、服务特定任务，而是具有广泛认知能力且能够胜任广泛任务，在大数据基础上进行训练并具有自我学习能力的模型。欧盟《人工智能法》经由“通用人工智能系统”过渡，以体系解释方法将生成式人工智能纳入“人工智能系统”范畴。为适应原本的风险分级布局，欧盟的做法是主要是通过在处理端引入“高影响能力”的模型自主性程度判断标准，〔48〕创设对应的“系统性风险”概念，并将具有此类风险的人工智能系统嵌入高风险规制范畴之中。

事实上，无需借助其他多余创设概念，而应当直接将通用人工智能模型纳入高风险情境。通用人工智能模型基于其系统本体能力属于较强的人工智能，存在更大可能突破预先设计的临界点，走向各类重大安全失控的巨大规模风险。〔49〕 这种潜在风险可能表现为模型数据训练引发的风险，如风险内容放大偏见效应，形成群体安全遭受不断破坏的恶性循环；可能表现为模型部署应用引发的风险，如生成涉及金融、国防、反恐等领域的虚假信息并扩大传播，实施直接威胁国家安全等重大安全的违法犯罪活动。而若照搬欧盟做法，创设其他概念嵌入原设的“高风险”分级体系，也可能会陷入更加严重的抽象定义不确定性困境中，比如“高影响能力”的概念设置就可能排除了小型模型造成巨大风险规模的情境。

四、高风险人工智能系统的规制原则与理念

在明确高风险人工智能系统之界定标准及主要情境的基础上，高风险人工智能系统的规制，应主要在特定情境中针对其安全性展开，在促进人工智能科技进步的同时，设置合理的规制原则与理念。

---

〔47〕 参见郑曦：“人工智能技术在司法裁判中的运用及规制”，《中外法学》2020年第3期，第680—681页。

〔48〕 参见陈亮、张翔：“欧盟生成式人工智能立法实践及镜鉴”，《法治研究》2024年第6期，第112—115页。

〔49〕 参见李伦主编：《人工智能与大数据伦理》，科学出版社2018年版，第265页。

(一)规制原则

国内当前对人工智能系统的规制原则的既有研究不少,〔50〕我国既有的两部有关人工智能法案的专家建议稿也都对规制原则进行了一定部署。本文认为,既有设计的原则数量过多,部分理念性的内容也不宜以“原则”加以涵盖,故可对规制原则体系进行重塑,将其归纳为合法性原则、科技伦理约束原则以及技术治理原则。

1.合法性原则

合法性原则的内容包括实体合法、程序合法两个层次。实体合法,在高风险人工智能系统运用于公共部门的情境中,首要是要坚持法律保留原则。持“侵害保留说”观点的学者认为法律保留范围不适用于给付行政,〔51〕但鉴于“高风险”对于重大安全状态的实质性显著减损明显涵盖了授益给付情境,故也应将其纳入属于法律保留范围的“重要事项”。〔52〕 换言之,对公权力应用人工智能系统进行辅助行权,可能对国家安全、社会公共安全、个人生命以及其他重要基本权利造成实质性显著减损的,都应得到法律的明确授权,否则构成违法。同时,为克服高风险人工智能系统的不完善之处,对辅助行权的人工智能决策也需实施有意义的人工监督,根据案件的是非曲直酌情作出公正的决策,保证决策实体合法,避免过度依赖人工智能系统。〔53〕

程序合法,在此表现为遵循程序正当原则下的公开原则,要求坚持系统的透明可解释、安全可问责。高风险人工智能系统的研发者、提供者、使用者应当依法以适当方式提供和说明人工智能产品和服务的基本信息、目的意图和主要运行机制等。为了满足透明原则的要求,人工智能主管部门应能够一直拥有访问训练数据、参数权重、程序和决策规则的权限,进而实现持续控制的义务,这些信息也应当留存并公示。与此同时,透明可解释原则也延伸要求,从事高风险人工智能系统提供活动的,应予以适当标注,不断提升标注的公平性、准确性和真实性。此外,在公共部门行权的高风险场境,程序合法也要求保障相对人享有受正当程序保护的权利,为其保留在面对人工智能不利决策时请求人工审查的意见表达渠道。相对人可通过行使事后的“人工智能纠正请求权”,请求行权的公共部门复查人工智能决策结果是否合法、合理。

2.科技伦理约束原则

科技伦理约束原则,宏观上是指发展人工智能应当坚持以人为本、智能向善,引导和规范

〔50〕 国内既有对人工智能规制原则的研究,主要呈现为列举式,涵盖了多元治理、包容审慎监管、数字福祉、以人为本等内容,参见郑智航:“人工智能算法的伦理危机与法律规制”,《法律科学》2021 年第 1 期,第 23—24 页;卢超:“包容审慎监管的行政法理与中国实践”,《中外法学》2024 年第 1 期,第 146 页;宋华琳:“人工智能立法中的规制结构设计”,《华东政法大学学报》2024 年第 5 期,第 8 页。但也有研究认为人工智能的规制原则展开不能停留在逐一罗列上,必须揭示不同原则之间的实质性关联,参见许可:“人工智能法律规制的第三条道路”,《法律科学》2025 年第 1 期,第 68 页。

〔51〕 参见林锡尧:《行政法要义》,元照出版有限公司 2006 年版,第 36—37 页。

〔52〕 参见马怀德主编:《行政法学》(第 3 版),中国政法大学出版社 2019 年版,第 49 页。

〔53〕 See Jorge Constantino, “Exploring Article 14 of the EU AI Proposal: Human in the Loop Challenges When Overseeing High-Risk AI Systems in Public Service Organisations,” *Amsterdam Law Forum*, Vol. 14, No. 3, 2022, p. 17.

人工智能产业健康有序发展。此项原则的内涵极其丰富，可通过作为高风险人工智能系统潜在致害对象的“重大安全”对其进行解构。首先，为保障国家安全、社会公共安全，其可被拆解出可持续发展原则、安全原则的意涵，指向相关主体应当采取必要措施保障所研发、提供和使用的高风险人工智能系统的可持续、安全。其次，为保障特定群体安全，可包括公平原则、保障弱势群体利益原则。公平原则表现为高风险人工智能系统的研发者、提供者、使用者，应当保护个人、组织的合法权益，不得实施不合理的差别对待，在公共部门应用情境，还需超越形式标准的实质上平等关照，通过组织程序和技术措施减少高风险人工智能系统的歧视。〔54〕 除遵循平等非歧视之外，高风险人工智能系统的研发者、提供者应当充分考虑未成年人、老年人、妇女、残疾人等弱势群体的权益保护。最后，为保护个人生命安全和其他重要基本权利安全，可从尊重人格尊严原则、可监督原则加以分解，确保人类能够始终监督和控制人工智能。

3.技术治理原则

“技术控制是风险治理机制的重要措施”，〔55〕有学者甚至直接提出，“规则代码化的技术主义规制进路和策略，已成为智能互联网时代不可阻挡的规制发展趋势”。〔56〕 对高风险人工智能系统的规制，除了深度融合法律规则与伦理规则的规范性治理，还应坚持技术治理，并将其上升到原则层面。高风险人工智能系统一般本体能力较强，可能造成重大安全的实质性显著减损，这些维度决定了实施治理的技术工具相匹配的高层次，建议由国家人工智能主管部门委托第三方科研机构开发出技术模型之后，推广至各级主管部门实施治理。技术治理原则的实施，应当贯通高风险人工智能系统“研发—提供—使用”的全生命周期，除了对高风险人工智能系统的工作机理进行探查、必要时调取相关数据等日常性的监督，还应重点体现在对高风险人工智能系统运行的训练数据、算法处理、内容输出等环节的限制上。比如当高风险人工智能系统的提供者未能及时采取措施阻止虚假有害信息等法律法规禁止的内容生成时，人工智能主管部门便可依靠技术治理工具，自动识别此类内容，作出停止其生成传输、消除等技术处置措施。

(二)包容审慎理念

包容审慎，是数字时代为破除传统监管法治困局而探索创新的新型监管理念。〔57〕 在人工智能风险规制中，包容审慎理念要求国家统筹发展与安全，在依法治理框架下给予高风险人工智能系统必要的创新发展时间与试错空间，并根据动态系统分析风险程度，适时、适度开展干预。

1.统筹发展与安全

发展和安全是辩证统一的，安全是发展的前提，发展是安全的保障。根据习近平法治思想的相关要求，当前中国的发展，既要善于运用发展成果夯实国家安全的实力基础，又要善于运

〔54〕 参见解志勇：“超级平台重要规则制定权的规制”，《清华法学》2024年第2期，第10页。

〔55〕 吴汉东：“人工智能时代的制度安排与法律规制”，《法律科学》2017年第5期，第135页。

〔56〕 马长山：“智能互联网时代的法律变革”，《法学研究》2018年第4期，第35页。

〔57〕 参见刘权：“数字经济视域下包容审慎监管的法治逻辑”，《法学研究》2022年第4期，第38页。

用法治方式塑造有利于经济社会发展的安全环境。〔58〕基于此,“统筹发展和安全”已然成为我国当前科技法治的重要目的和原则,对我国人工智能立法具有根本意义。故对高风险人工智能系统的规制,也应坚持统筹发展与安全的理念,不宜以单纯的风险防范为单一的立法目的,而应以促进高风险系统研发应用与防范主要风险为二元立法目的,“实现高质量发展和高水平安全良性互动”。〔59〕

2.结合鼓励创新与依法减负

包容审慎理念,从传统规制理论角度观察,与回应性规制提供更加完备且多元的执法策略存在契合之处,也即要求规制应当将鼓励创新与传统的依法治理相结合。创新是人工智能法律的价值灵魂,鼓励创新体现在创新促进规范的正面激励与减负支持两个层面。人工智能立法一方面应当细化规定正面激励制度,推进自动驾驶、诊疗服务等主要情境技术的研发和应用;完善知识产权创新激励机制,通过权利保护、交易和限制等制度,促进技术创新与产业发展。另一方面,人工智能立法及相关政策制定也应当提供减负支持,包括减少合规成本、税收扶助等,比如对从事人工智能基础研究进行专项资金支持,对人工智能重点项目运营企业给予税收优惠等。

3.动态系统分析弹性风险

威尔伯格的动态分析论认为,不同作用力各自的汇合与各自的强弱程度并不绝对或固定,它们在特定案件的特定图景中形成的动态合力才起决定性作用。〔60〕高风险人工智能系统造成的具体风险合力,原则上只有在满足上文的三大界定维度,同时应用于本文列举的九类情境之时,才可归类为高风险。但如果某一要素以特殊的强度发生作用,它也可能会成为“高风险”认定正当化的理由。在不同要素之间,比如某一人工智能系统运用的情境可能涉及规模巨大的不特定多数人的财产性利益,此时其不一定受限于“重要基本权利”要素多用于人身性权利的框架,需要综合考虑该人工智能系统的涉众性程度、使用者财产权利受损风险与收益的差值等弹性指标,最终决定对其的规制严格程度。而在同一要素之中,比如受损维度之下的风险规模指标,系统本体规模与系统作用对象规模两类指标,若有一类特别强大,也可满足风险规模“巨大”的认定要素。

当然,与这种动态系统分析相伴随的可能是规制方式的动态化,这可能与法安定性相背离。因此,按照信赖利益保护原则,规制主体在必要时需对受影响的相关方作合理补偿。〔61〕

(三)合作规制理念

合作规制理念,要求在高风险人工智能系统的规制中,推动形成规制与监管共同发力的局

〔58〕 参见《习近平法治思想概论》编写组:《习近平法治思想概论》,高等教育出版社 2021 年版,第 282 页。

〔59〕 参见“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(2024 年 7 月 18 日中国共产党第二十届中央委员会第三次全体会议通过)”,载《人民日报》2024 年 7 月 22 日,第 3 版。

〔60〕 周晓晨:“过失相抵制度的重构——动态系统论的研究路径”,《清华法学》2016 年第 4 期,第 111 页。

〔61〕 参见王贵松:“风险行政与基本权利的动态保护”,《法商研究》2022 年第 4 期,第 31 页。

面，尝试开放互动的实验主义治理模式，强调用户协作的安全风险防控机制。

1.推动形成规制与监管共同发力局面

推动形成规制与监管共同发力的局面，并不是说“规制”与“监管”全然分立，事实上两者不是非此即彼。“监管”是较为狭义的“规制”，主要是指行政机关或法律授权的机构基于规制职责对市场主体的准入、经营管理和退出等实施的监督管理。〔62〕“规制”的外延相较“监管”会更加广阔，且其强调的多元主体、灵活方式与多样手段也比“监管”侧重的单向、强硬监督管理关系，更为契合人工智能系统敏捷治理的趋势。毕竟在人工智能规制中，监管主体不可能时刻对每个被规制者进行监督，〔63〕自我规制的理念根基也在于此。因此，在对高风险人工智能系统进行规制时，除了需要坚持自上而下的政府监管，还应结合企业自我规制、第三方规制等方式，科学布局多元共治，以信任、透明、可问责的制度文化应对快速变化的风险。〔64〕

2.推行开放互动的实验主义治理模式

合作规制除鼓励自我规制、第三方规制，还可通过契约关系构建容错机制，建立高风险人工智能系统的实验主义治理模式，尽可能消除实验参与方的后顾之忧。这种治理模式的一种表现形式，就是欧盟《人工智能法》第六章规定的监管沙盒制度。监管沙盒制度提供了一个试验人工智能技术或产品的封闭安全法律政策空间，让监管机关可以直接了解创新技术或产品的设计及其发展，从而更好地调整现有法律法规和引入新的法律。对于人工智能系统而言，在风险定性尚未确定的情境中，实验主义治理模式能够通过设置高风险定级的原始实验性框架维度，为潜在的高风险新技术留足发展空间，也能够赋予实验参与者更大的自主权，鼓励其参与检测不同监管方案的效果，进而实现一种“有条件、有限度、有控制的放松监管”。〔65〕

3.强调用户协作的安全风险防控机制

合作规制的理念也要求在高风险人工智能系统的安全风险防控机制建设中强调用户协作，这里的“用户”指向的是人工智能的使用者。如果说研发者、提供者主要是从前端控制人工智能风险，那么使用者则主要影响的是人工智能产品和服务的后端，是风险转化为现实损害的“最后一公里”。比如，用户能够决定什么时候启用自动驾驶功能、从输入端赋予生成式人工智能大模型哪些关键词等，这些决定都能够直接影响人工智能的风险转化与风险定级。因此，在部分情境下，相较于研发者、提供者，使用者对于高风险人工智能系统具有更为直接的控制力，其也应当分担一定的预防和管理人工智能风险的义务，比如控制输入数据质量的义务、更新和维护人工智能系统的义务等。还有部分情境，使用者则需对人工智能负担监督义务，比如涉及诊疗服务情境的事后再判断义务，涉及自动驾驶情境中的紧急接管义务等。

---

〔62〕 参见马英娟：“监管的语义辨析”，《法学杂志》2005年第5期，第113页。

〔63〕 参见(英)罗伯特·鲍德温、马丁·凯夫、马丁·洛奇：《牛津规制手册》，宋华琳、李鸽、安永康、卢超译，宋华琳校，上海三联书店2017年版，第183页。

〔64〕 See Keith E. Sonderling and Bradford J. Kelley, “Filling the Void: Artificial Intelligence and Private Initiatives,” *North Carolina Journal of Law & Technology*, Vol. 24, No. 4, 2023, p. 157.

〔65〕 廖凡：“论金融科技的包容审慎监管”，《中外法学》2019年第3期，第810页。

五、高风险人工智能系统的情境化规制展开

高风险人工智能系统的情境化规制，应当以重大安全维护为规制目标，在我国未来的专门立法进路上采取先纵后横、试验性立法的模式，最终达致制定《人工智能安全法》的目标。

（一）规制立法进路

当前我国已初步具备了人工智能专门立法的一定基础，仍需明确的是规制立法的具体路径、规制模式、规制单元、规制目标。

1. 先纵后横的立法路径

我国已初步形成了以网络信息等领域立法为主体，以面向“生成式人工智能服务”“深度合成”等问题的部门规章、地方性法规等作为延伸的纵向治理规则框架，但还缺少横向专门的人工智能法律。为推进 2025 年初步建立人工智能法律法规，到 2030 年建成更加完善的人工智能法律法规立法规划的实施，〔66〕应当加快立法步伐，在先行的纵向“小快灵”基础上形成未来的横向“大部头”立法，为高风险人工智能系统的分级定性提供更高层次的规范依据。同时，在法律之外，还应注重人工智能技术标准和伦理规范的制定，增强整体规范体系的协调性。〔67〕

2. “行为规制＋个体赋权”

从理论上讲，立法对人工智能风险进行规制，可以采取两种不同的规制模式。一种是与人工智能技术特征与风险特征高度契合的行为规制模式，即对人工智能的研发、提供、使用，以及上市后的维护、更新、召回等贯穿人工智能系统全生命周期的行为予以系统性规制，通过对研发、提供、使用者设定行为规范、施加法律义务的方式来实现控制风险的目的。〔68〕另一种模式是赋权模式，即赋予可能受到人工智能风险影响的主体以一定的权利，通过权利人向义务人主张权利的方式，控制人工智能风险。这两种模式虽然强调保护的侧重点不同，但并不互相排斥，甚至可以同时出现在一部立法之中。鉴于本文界定的高风险人工智能系统，其潜在的威胁对象既包括国家安全、社会公共安全等集体性风险，也包括个人生命安全、其他重要基本权利安全等个体性风险，故宜在采取行为规制的同时辅之以个体赋权的方式进行规制。如此，既可有效弥补政府监管力量的不足，也能弥补以抽象要件标准判断风险分级的局限性。〔69〕

3. 以情境作为高风险规制单元

2022 年，中共中央办公厅、国务院办公厅印发《关于加强科技伦理治理的意见》，将“敏捷治理”列为五大治理要求之一。2023 年，国家互联网信息办公室发布的《全球人工智能治理倡议》，也倡议“实施敏捷治理，分类分级管理，快速有效响应”。敏捷治理强调尊重人工智能发展规律并保持跟踪研判，强调治理节奏上的快速回应和尽早介入，在治理规则上推进弹性原则与

〔66〕 参见《国务院关于印发新一代人工智能发展规划的通知》（国发〔2017〕35 号）。

〔67〕 参见宋华琳，见前注〔50〕，第 20 页。

〔68〕 See Hannah Ruschemeier, “AI as a Challenge for Legal Regulation-The Scope of Application of the Artificial Intelligence Act Proposal,” *European Research Area Forum*, Vol. 23, No. 3, 2023, p. 364.

〔69〕 参见周学峰：“论人工智能的风险规制”，《比较法研究》2024 年第 6 期，第 51—52 页。

具体类型化规则的有效结合，已然成为我国人工智能治理中的一项重要策略。而以情境作为高风险人工智能系统的规制单元，既可提升规制的动态性，也能够以可变动的情境增强立法的适应性，与上述敏捷治理的规制策略相契合。相反，若单纯依赖于抽象的行为规范、艰涩的技术标准加以规制，缺失特定情境下的合法性判断、价值理念以及利益权衡，高风险的识别与认定会陷入无的放矢的困局。

4.以安全作为高风险规制目标

在人工智能时代，安全也是人工智能规制立法的核心价值。〔70〕对高风险人工智能系统的情境化规制措施，其规制目标就是用于化解高风险人工智能系统可能对“重大安全”造成实质性显著减损的风险。这种安全目标要求人工智能全过程的安全保障体系和能力建设，对公共部门提出了保障性义务的命题。对应用于公共部门的情境，公共部门负有对高风险人工智能系统的持续监督义务，以及确保系统合法有效运行的组织程序义务；在高风险人工智能系统失灵时，公共部门还负有介入和最终决策的义务。同时，在全球视野下，安全目标的实现还需要在域外立法上配备反制其他国家对我国人工智能发展的遏压的规则，〔71〕打造高风险人工智能系统发展的全球安全环境。

（二）事前规制

对高风险人工智能风险进行事先规制，主要通过管理情境化具体清单、制定高风险技术标准、设置安全评估以及预防型备案制的方式开展，彰显“智能法律预测化”导向。〔72〕

1.情境化具体清单管理

高风险人工智能系统的情境化清单管理，首先需要确定此管理系统可能威胁国家安全、社会公共安全、个人生命安全、其他重要基本权利等重大安全领域，其次这种威胁应当对重大安全构成实质性“显著减损”程度的威胁。在此基础上，本文明列的具体清单，涵盖了涉及关键数字基础设施、生物识别、教育就业、自动驾驶、诊疗服务、新闻服务、授益给付、司法程序、通用人工智能共九类情境。此类情境化的具体清单管理，应当留予一定的可变动性口径，如果出现新的特定人工智能系统应用情境满足原定维度，规制者有权通过授权立法的方式，将此类系统增加纳入情境清单之中，反之，规制者也有权通过授权立法的方式将不再符合原定维度的人工智能系统剔除。

2.高风险技术标准制定

标准延伸了法律的规范作用，使得以权利义务配置为内容的抽象的法律规范之规范性落到实处；而认证制度与标准的配合，则是通过事前的“合格评估”来促使生产者遵守标准，从而确保标准的有效实施。〔73〕对于高风险人工智能系统的事前规制，仅依靠“高风险”的抽象内涵构成要件是不足以精准定位规制对象的，仍需就具体的“显著减损程度”的标准制定技术标

〔70〕 参见吴汉东，见前注〔55〕，第134页。

〔71〕 参见张吉豫：“赋能型人工智能治理的理念确立与机制构建”，《中国法学》2024年第5期，第71页。

〔72〕 余成峰：“法律人工智能新范式：封闭与开放的二元兼容”，《中外法学》2024年第3期，第599页。

〔73〕 参见柳经纬：“标准与法律的融合”，《政法论坛》2016年第6期，第26—28页。

准，尤其是对人工智能基础模型的算力、参数、使用规模等设置一定的技术临界值，以便在动态系统分析中更为直观、准确地得出是否构成“高风险”的判断结论。同时，高风险人工智能系统的标准内容涉及“重大安全”，属于《中华人民共和国标准化法》第 10 条规定的情形，应由人工智能主管部门就“高风险”制定强制性的国家技术标准，以国家力量确立人工智能的能力评测、应用评定体系。

3. 自评估或第三方评估

基于高风险人工智能系统的复杂性和与之相关的风险，需要针对高风险人工智能系统设定安全风险评估制度。具体而言，此类人工智能系统的研发者、提供者应当在提供产品和服务前开展安全风险评估，并且对处理情况进行记录，在提供活动过程中也应保持至少每年一次的定期评估频率。评估可由自设的内部机构开展，也可以委托第三方机构，但评估机构均应当符合具有独立性、资质能力、无利益冲突和适当的网络安全要求等条件。对于关键数字基础设施或者人工智能基础模型的算力、参数、使用规模特别巨大的高风险大模型，为避免造成难以承受的安全减损后果，建议由国家人工智能主管部门进行额外的定期抽查评估。

评估的具体内容，原则上应当包括是否存在潜在的偏见或者歧视，对“重大安全”的实质影响及显著减损程度，是否依法开展科技伦理审查，保护措施是否合法、有效并且与安全减损程度相适应等。若评估系统所涉情境为群体性安全、个人生命安全或其他重要基本权利安全的，尤其是用于授益给付情境时，评估还需重点考量在特定情境下系统的预期目的、系统的使用期限和频率、可能影响的自然人和群体的类别、产生的具体危害风险、人工监督与治理措施设置等内容。

4. 通用人工智能备案制

有观点将规范意义上的行政备案划分为预防型备案、告知型备案与后设型备案三种类型。〔74〕本文借用此框架，认为对于涉及通用人工智能模型的高风险人工智能系统，宜采取预防型备案这一事前备案形式，以防范特定情境下的安全风险为核心目标，要求此类模型提供者在从事人工智能特定提供活动前应当通过国家人工智能监管平台进行备案并履行相关手续。这种事前备案，虽然只是一种形式审查，但也不失为一种实现对相关模型研发、提供活动的积极、事前监管。

（三）事中规制

对高风险人工智能系统的事中规制，主要依靠强化此类系统研发、提供、使用者的义务体系建设，健全风险监测预警、报告响应机制，建立科技伦理跟踪审查工作机制等手段。

1. 研发、提供、使用者义务体系建设

非产品类的人工智能是欧盟《人工智能法》的主要规制对象，其主要针对提供者、部署者、进口商和分销商四类主体，而大量的法律义务重点聚焦于提供者和部署者，其中人工智能提供者还可能因人工智能价值链上的责任规定（名称或商标标注、重大修改、改变预期目的等）而扩

〔74〕 参见王由海：“行政备案的实践类型与法治化路径”，《法商研究》2023 年第 1 期，第 78—79 页。

张其范畴;〔75〕人工智能系统部署者,即在其授权下使用人工智能系统的主体。〔76〕不难发现,欧盟是将纯粹的人工智能研发行为排除在规制之外的,这与我国当前就生成式人工智能服务规制对象排除未面向公众提供服务的人工智能研发、应用行为是一致的,〔77〕因为此类情境不具有风险的可实现性。而实质上,欧盟采用的所谓"提供者"的意涵是广义的,其实包含了研发者和负责狭义的"提供投放运营"的提供者角色,部署者则是指事实上的使用者。是故,在我国语境下,使用"研发者""提供者""使用者"的称谓更易理解,也能与国内既有的纵向规则框架相匹配。

具体而言,首先,提供者的义务包括风险管理义务、数据治理义务、技术可靠义务、环境保护义务、信息提供义务、质量管理义务、模型登记义务等。在涉及通用人工智能模型的情境中,提供者还应额外遵循透明度义务、防止生成非法内容义务、披露受版权保护的训练数据义务。其次,在配合提供者义务的同时,研发者的独立义务包括在产品设计中嵌入非法使用限制的义务、标注使用数据来源义务、提升数据训练质量义务等。最后,使用者的独立义务,则主要包括禁止利用人工智能从事非法活动的义务、人工智能生成内容知识产权认定的披露义务等。

2.风险监测预警、报告响应机制

高风险人工智能系统的正常使用,也存在对"重大安全"造成实质性显著减损的可能,但当其尚未构成迫切威胁时,属于一种远期风险。对此类风险的事中规制,首先是常规的风险监测,这种监测既包括此类系统的研发、提供者有义务向国家人工智能主管部门定期报告系统运转情况,也包括系统研发者、提供者自行建立的人工智能安全风险披露机制。其次,应以科学不确定性原理为核心依据,在国家层面构建高风险人工智能的风险预警机制。〔78〕依据"科学不确定性原理",只要有能被确认的重大安全受损的风险迹象存在,便可作为预警机制启动的条件。作为预警信息的收集、报送、发布主体,系统研发者、提供者应当把握时间窗口,及时发布自主性预警,防患于未然或避免损害扩大。而相关主管部门作为收到报送预警信息方,也应在判断后决定是否发布行政性预警,由此形成预警主体的分工协同结构。〔79〕当威胁彻底造成实质损害,发生人工智能安全事件时,则需启动报告响应机制。系统研发者、提供者应采取补救措施,包括在必要时采取中断运行的紧急措施,并及时向人工智能主管部门报告。主管部门若认为特定系统不符合规制要求,也可以要求研发者、提供者采取一切适当的纠正措施以使人工智能系统符合要求,包括从市场上撤回人工智能系统,或在其可能规定的期限内召回人工智能系统等。

3.科技伦理跟踪审查工作机制

高风险人工智能系统的研发者、提供者在事前规制的自评估或第三方评估中,就包括评估

〔75〕 参见欧盟《人工智能法》第16条、第25条。

〔76〕 参见欧盟《人工智能法》第26条。

〔77〕 参见《生成式人工智能服务管理暂行办法》第2条。

〔78〕 参见解志勇:"公共卫生预警原则和机制建构研究",《中国法学》2021年第5期,第232—233页。

〔79〕 参见韩大元、莫于川主编:《应急法制论——突发事件应对机制的法律问题研究》,法律出版社2005年版,第228页。

是否依法开展科技伦理审查的评估内容。而在事中规制阶段,需进一步强化科技伦理跟踪审查工作机制建设,科技伦理审查委员会应持续跟踪监督高风险人工智能系统研发、提供、使用活动的全过程。在跟踪审查中,坚持以人为本、智能向善,贯彻对人的尊重、对人权的保护,对系统运用的负面影响进行防范和制止,〔80〕抑制不公正、不道德因素对于系统的渗入。〔81〕跟踪审查的主要内容,主要包括系统实施方案执行情况及调整情况、科技伦理风险防控措施执行情况、科技伦理风险的潜在变化及可能影响研究参与者权益和其他重大安全等情况等,必要时也可作出暂停或终止活动等决定。对属于具体清单管理范围内的高风险情境,尤其需要加强跟踪审查和动态管理,跟踪审查的间隔也应短于低风险系统。

(四)事后规制

对高风险人工智能系统的事后规制,在行政法律责任承担方面,主要包括行政处罚、失职问责以及行政救济等形式。

1.行政处罚

对于高风险人工智能系统适用于私营的情境,研发、提供、使用者违反法律义务,一般由人工智能主管部门给予包括警告、没收违法所得、责令暂停或停业整顿相关业务、吊销相关业务许可或者吊销营业执照等决定在内的行政处罚决定。此类行政处罚决定的作出,应结合风险动态,充分考虑包括违法行为的严重程度、违法主体主观状态、补救措施、违法者具体规模等各种因素,采取合乎比例的惩戒措施。若此类系统的研发、提供、使用者建立了内部合规计划,且该合规计划具备预防违法犯罪行为发生的“有效性”并得到切实执行,〔82〕行政机关也可在其主动配合调查、整改、认罚的情形下,考虑作出减免处罚的决定。

2.失职问责

对于应用于辅助公共部门行权情境的高风险人工智能系统,若此类系统的使用失范,对国家机关等公权力主体不宜适用行政处罚,此时一般应由其上级机关或者有关人工智能主管部门责令改正。对直接负责使用此类系统的主管人员和其他直接责任人员,依法给予行政处分。而若依法不构成行政执法过错的情形,则不再追究有关工作人员的行政执法责任。

3.行政救济

此外,若特定高风险人工智能系统的研发、提供活动属于法律、行政法规规定应当取得许可的情形,也可能存在研发、提供者作为行政相对人,对行政机关作出的行政许可决定(包括不予许可、中止许可乃至撤销许可决定等)不服的情况。此时,研发、提供者也可依法申请行政复议或者向人民法院提起行政诉讼。

六、结 语

坚持人工智能风险分级规制的思路,并不应该因为风险评估困难、分类可能疏漏的客观困

〔80〕 参见张文显:“构建智能社会的法律秩序”,《东方法学》2020 年第 5 期,第 17 页。

〔81〕 参见马长山:“人工智能的社会风险及其法律规制”,《法律科学》2018 年第 6 期,第 53 页。

〔82〕 参见解志勇、那扬:“有效企业合规计划之构建研究”,《法学评论》2022 年第 5 期,第 165 页。

难就踌躇不前。面对人工智能风险，主要规制对象为“高风险”人工智能系统，在“风险＋情境融合治理”的思路指引下，提出更具针对性和实操性的界定维度和标准，并保持“高风险”系统应用情境的开放性和可变动性。将“重大安全显著减损”作为“高风险”的界定标识，通过明列“重大安全”的涵盖类型、“显著减损”的影响要素，以动态系统分析理论辅助完成高风险的分析与认定难题，为铺开包容审慎、合作规制提供精准定位。未来，我国人工智能立法，在高风险人工智能系统的规制层面，也应坚持以情境作为主要规制单元，以重大安全维护作为规制目标，在促进人工智能科技进步的同时，筑牢全过程规制防线。

Abstract: The underlying scientific logic of artificial intelligence technology determines that its risks are cognizable, describable, analyzable and definable. Unlike the general risk governance established by the European Union's Artificial Intelligence Act, China can adopt the approach of "risk + contextual integration governance" to respond to the systemic risks of High-risk artificial intelligence with high quality. The so-called "high-risk" definition standard can be recognized based on three dimensions, such as the strength of the AI ontology, the function of the object, and the degree of potential harm. Firstly, the risk level is positively correlated with the strength of the AI body, and both strong AI and superintelligent AI are high-risk systems. Secondly, the direct object of its function is "significant security", covering national security, public security, personal life security and other important basic rights security. Thirdly, there is a possibility of substantial impairment of "significant security". In view of the above standard, the governance and regulation of high-risk AI is mainly focused on its safety, and should follow the principle of legality, the principle of scientific and technological ethical constraints, the principle of technological governance, and uphold the concept of inclusive prudence and cooperative regulation. Taking the situation as the governance unit, with safety maintenance as the primary goal, and adopting the legislative model of behavioral regulation combined with individual empowerment, the whole process of governance is carried out before, during and after the event. Ultimately, the legislative goal is to enact the "Artificial Intelligence Security Law".

Key Words: Artificial Intelligence; High-risk Artificial Intelligence; Artificial Intelligence Security Law; Governance; Regulation

(责任编辑:彭　錞)