从单向监管到协同共治：生成式人工智能风险治理的范式转换

梁静[1]，周伟萌[2,*]
1. 玉林师范学院 历史与法律学院，广西玉林 537000
2. 玉林师范学院 发展规划处，广西玉林 537000

摘要：生成式人工智能的快速迭代正在重塑信息生产格局，也带来了风险形态的深刻变迁。面对生成式人工智能风险的系统性、不确定性与分布式等特征，以政府为核心的传统单向监管模式逐渐陷入治理困境。因此，有必要推动生成式人工智能风险治理从单向监管向协同共治的范式转换。协同共治范式强调主体多元化、规范协同化、机制敏捷化与责任共担化。构建协同共治治理体系，需要明确政府、平台企业、行业组织、第三方机构与用户等多元主体的角色定位，建立风险沟通、分类分级、软硬法衔接、全链条协同与敏捷监管等核心运行机制，并从完善法律制度、健全协同机制、强化技术赋能三个层面系统推进。

关键词：生成式人工智能；协同共治；多元主体；敏捷治理

From Unilateral Regulation to Collaborative Governance: Paradigm Shift of Generative AI Risk Governance

Jing Liang[1], Weimeng Zhou[2,*]
1. Yulin Normal University, School of History and Law, Yulin, Guangxi, 537000
2. Yulin Normal University, Development Planning Office, Yulin, Guangxi, 537000

Abstract: The rapid iteration of generative AI is reshaping the landscape of information production and bringing about profound changes in risk patterns. Faced with the systematic, uncertain, and distributed characteristics of generative AI risks, the traditional one-way regulatory model centered on the government has gradually fallen into a governance dilemma. In this regard, it is necessary to promote a paradigm shift in generative AI risk governance from unilateral regulation to collaborative co-governance. The collaborative co-governance paradigm emphasizes pluralistic subjects, normative synergy, agile mechanisms, and shared responsibility. Building such a governance system requires clarifying the roles and positioning of multiple subjects including the government, platform enterprises, industry organizations, third-party institutions, and users; establishing core operational mechanisms such as risk communication, classification and grading, hard-soft law coordination, whole-chain collaboration, and agile regulation; and advancing systematically through three dimensions: improving legal institutions, strengthening collaborative mechanisms, and enhancing technological empowerment.

Keywords: Generative AI; Collaborative Governance; Pluralistic Subjects; Agile Governance

以 ChatGPT、DeepSeek 等大语言模型为代表的生成式人工智能，使社会经济发展进入到智能生产的新阶段。这类技术在内容创作、代码生成与知识服务等领域展现出了前所未有的惊人能力，极大降低了信息生产的门槛与成本。但生成式人工智能在赋能生产的同时，也不断衍生出了虚假信息扩散、知识产权侵权、算法偏见、数据安全隐患及深度伪造等多重风险，这就对个人权益、公共利

益乃至国家安全形成新的挑战。我国目前对生成式人工智能的治理仍以行政监管为主，《生成式人工智能服务管理暂行办法》[1]较快搭建了风险管控的制度框架，但在面对生成式人工智能技术迭代快、应用场景广、风险传导链条长等新特点时，单一依靠政府的监管模式逐渐力不从心。学界对此已展开较多讨论，但现有研究主要还是关注具体制度的设计或单一工具的改进，对生成式人工智能风险治理的范式尚缺乏系统的理论阐释与整体建构。生成式人工智能风险的系统性、不确定性与分布式特征，从根本上决定了单靠政府一方难以实现有效治理，必须推动治理范式从单向监管转向多元主体协同共治。基于此，从生成式人工智能风险的新特质切入，剖析单向监管模式的治理困境，阐释协同共治的理论逻辑与核心内涵，构建多元主体协同的运行机制，提出我国推进协同共治的路径。

1 生成式人工智能风险新特质与单向监管困境

1.1 生成式人工智能风险新特质

相较于传统人工智能，生成式人工智能在风险形态呈现出新特质。第一，风险的系统性与传导性。生成式人工智能的风险贯穿数据输入、模型训练、内容输出到用户反馈的全链条，各环节风险相互交织，特别是输入端的数据合规问题可能经过模型训练放大为输出端的内容风险，造成闭环式累积；第二，风险的不确定性与涌现性。大模型的黑箱特性使得开发者难以完全预判其行为。随着模型能力边界拓展，还可能涌现出超出设计预料的能力与风险，这类涌现性风险很难通过事前评估完全识别；[2]第三，风险的分布式与多主体性。生成式人工智能的价值链条包括数据提供者、模型开发者、平台运营者、应用开发者及终端用户等多个主体，风险存在于这个价值链的各个节点；第四，风险的扩散性与规模化。生成式人工智能极大降低了内容生产的门槛，使得虚假信息和有害内容的生产传播趋于自动化、规模化，风险扩散速度呈指数级增长，传统监管手段在响应速度上难以跟上。截至 2026 年 2 月，全球探测到超过 23 万例 OpenClaw 公网暴露实例，其中约 8.78 万例存在数据泄露，约 4.3 万例存在个人身份信息暴露。[3]

1.2 单向监管模式治理失灵

所谓单向监管，是指以政府为唯一监管主体、以命令控制为核心手段、自上而下实施规制的治理模式。其特征主要体现在四个方面：（1）主体单一，政府包揽规则制定、执行监督与违法处罚等全部职能，其他社会主体缺乏参与治理的制度化渠道；（2）方式刚性，主要依赖法律规范与行政命令，通过事前审批、事中检查、事后处罚实现监管目标，规则调整周期长，难以随技术发展动态调适[4]；（3）逻辑静态，基本预设是风险可预测、规则可预设、结果可控制，但生成式人工智能的不确定性与涌现性恰恰打破了这一预设；（4）结构垂直，各监管部门按行业和领域划分权限，条块分割明显，应对跨领域综合风险时容易出现监管重叠与真空并存的局面。单向监管的这些内在特征与生成式人工智能风险新特质之间存在矛盾，从而导致一定程度的治理失灵，主要体现在三个方面：（1）认知局限带来监管偏差。监管者与被监管者之间信息不对称严重，监管部门技术储备不足，难以对模型风险作出准确判断；（2）效率困境导致监管滞后。生成式人工智能技术迭代极快，而单向监管下规则的制定与调整需经过复杂程序，监管回应速度远远落后于技术发展速度，始终处于被动追赶状态；（3）覆盖盲区形成监管漏洞。单向监管以平台企业为主要对象，侧重于应用层面的内容管控，对数据来源、模型训练、开源社区等上游环节的监管相对薄弱，对开源模型、API 接口调用等新兴业态更是难以有效覆盖[5]。

2 从单向监管到协同共治：范式转换的理论逻辑

2.1 协同共治的理论基础

生成式人工智能风险治理从单向监管转向协同共治，有着深厚的理论支撑。首先，传统管理思维以政府为单一中心的治理模式已难以应对日益复杂的公共事务，主体多元化已经成为治理理论最核心的主张。公共治理的主体应当是多元的，政府、市场与社会组织各有其角色，治理的实质在于基于市场原则、公共利益与认同之上的合作；其次，德国法社会学家贡塔·托依布纳的反身法理论为规制方式转型提供了新视角，该理论认为现代社会功能分化导致各子系统形成自身独特的运行逻辑，法律系统的认知能力有限，无法直接深入其他子系统内部进行精细化规制，法律应当放弃直接干预，转而通过影响其内部结构与运行程序来推动自我规制；最后，动态治理理论也为机制创新提供了方法论指引，面对人工智能的不确定性，法律不应试图通过静态规则体系实现一劳永逸的治理，而应构建多元主体基于沟通的动态治理机制，围绕风险对称的民主性沟通来控制不确定性，通过沟通程序的制度设计确保多元利益相关者在信息对称的条件下敏捷作出具有可问责性的风险决策[6]。

2.2 范式转换的理论逻辑

从单向监管转向协同共治，是生成式人工智能风险治理的客观要求。从技术与效率维度看，生成式人工智能风险的系统性、不确定性与分布式特征，决定了任何单一主体都难以仅凭自身力量实现全面有效的治理；而企业自我规制、行业自律管理与社会公众监督参与，能够在政府监管之前形成多层次的风险过滤机制，显著降低治理成本、提升治理效率。从价值与规范维度看，协同共治有利于实现发展与安全的价值平衡，通过多元主体共同参与，能够在治理过程中充分吸纳各方利益诉求与价值主张，在有效防控风险的同时为技术创新保留必要空间；同时，人工智能技术的快速迭代与法律规范的相对稳定之间始终存在时间差，由此产生的规范延迟现象使得单一的静态规则体系难以应对持续演变的技术风险，必须通过多元主体的协同互动与规则的动态调整来弥合规范与现实之间的差距。[7]从全球维度看，协同共治是国际人工智能治理的共同趋势，推动我国治理向协同共治转型，既是顺应全球治理趋势的需要，也是提升我国在国际人工智能治理中话语权的必然要求。

2.3 协同共治范式的核心内涵

生成式人工智能风险的协同共治范式，指的是政府、平台企业、行业组织、第三方机构与用户等多元主体基于共同治理目标，通过制度化的沟通协作机制，各负其责、协同行动，共同应对生成式人工智能风险的治理模式。其核心内涵体现在四个维度：一是主体多元化，打破政府作为唯一监管主体的传统格局，各主体基于各自资源禀赋与能力优势承担不同治理角色，形成既相互协作又相互制约的治理网络[8]；二是规范协同化，强调硬法规范与软法规范的有机结合与动态衔接，硬法设定底线要求与基本框架，软法包括技术标准、行业准则、伦理规范等，二者协同配合既能保障治理权威性，又能提升治理适应性；三是机制敏捷化，以敏捷治理为基本取向，强调治理机制对技术变化与风险演变的快速响应能力，通过分类分级管理、风险动态评估、监管沙盒试点等实现治理措施与风险水平的动态匹配；四是责任共担化，治理责任不再由政府独担，而是由多元主体按各自角色定位共同分担，平台企业承担主体责任，行业组织履行自律责任，第三方机构提供专业支撑，用户承担合理使用与监督责任，政府则承担统筹协调与最终监管的责任。

3 协同共治的主体架构与运行机制

3.1 多元主体的角色定位

构建生成式人工智能风险协同共治体系，首先，要明确各治理主体的角色定位与责任边界。政府从全能监管者转向元治理者，承担规则制定、统筹协调与监督兜底职能，负责制定基本法律框架

与底线规则，推动建立协同治理机制，将具体治理事项交由市场与社会主体承担；其次，平台企业要从被动的被监管者转向主动的治理参与者。作为服务的直接提供者和风险治理的第一道防线，承担建立内部安全治理体系、开展风险评估与安全测试、配合政府监管与行业自律等主体责任。内容传播平台和应用程序分发平台有义务核验相关标识材料。2025 年 3 月，网信办联合工信部、公安部等发布《人工智能生成合成内容标识办法》，搭配强制性国家标准。[9]规范要求 AIGC 服务提供者对生成合成内容添加显式标识（让公众能辨别），并对文件元数据添加隐式标识（用于溯源监督）；再次，行业组织要从边缘参与者转向自律管理者。要发挥其连接政府与企业的桥梁作用，负责制定行业标准与自律规范、开展行业自律检查，其专业优势在技术标准制定、最佳实践推广等方面能够发挥政府难以替代的作用；另外，第三方机构要从缺位走向专业支撑。其包括安全评估机构、算法审计机构、检测认证机构、科研院所等，负责提供安全评估、算法审计、技术检测等专业服务，其独立性与专业性能够有效弥补政府与企业在技术能力方面的不足；最后，用户要从治理对象转向治理参与者。其既是服务的使用者，也是风险治理的重要参与者与监督者，应承担合理使用、监督举报、参与治理规则形成等角色，其体验反馈与行为选择对治理效果具有重要影响[10]。

3.2 协同共治的核心运行机制

主体架构的搭建只是基础，要使多元主体真正形成治理合力，还需要建立健全一系列核心运行机制。第一，风险沟通机制，其是协同共治的基础性机制，也是实现动态治理的关键。搭建风险信息共享平台，能实现政府、企业、行业组织与第三方机构之间必要的风险信息互联互通；第二，分类分级机制，其是实现精准治理与敏捷治理的基础制度。分类分级是综合考虑应用领域、功能类型、影响范围、危害程度等因素划分风险等级，根据风险等级匹配相应的治理要求，并建立风险等级的动态调整机制，实现治理强度与风险水平的动态匹配[11]；第三，软硬法衔接机制，即明确硬法与软法的功能定位与边界划分。建立软法的制定与认可机制，推动硬法与软法的动态衔接，硬法的原则性规定通过软法加以细化落实，软法的成熟经验通过立法上升为硬法规范；第四，全链条协同机制，即明确研发、部署、应用、反馈各环节的治理重点与责任主体，建立各环节之间的信息传递与风险传导机制，推动价值链各主体之间的协同治理，形成风险联防联控的链条；第五，敏捷监管工具机制，即强调借助监管沙盒、合规试点、快速响应与规则迭代等多样化工具，提升治理的灵活性与适应性，在有效管控风险的同时为技术创新留出空间。

4 我国生成式人工智能协同共治的具体路径

4.1 完善协同共治的法律框架

首先，明确各主体的法律地位与权责边界，在人工智能专门立法中系统规定各治理主体的法律地位、权利义务与责任形式。特别是明确平台企业的主体责任范围、行业组织的自律管理权限、第三方机构的准入条件与执业准则，确立用户参与治理的渠道与方式；其次，建立分类分级的动态监管制度，在总结现行监管实践经验的基础上，建立科学系统的生成式人工智能分类分级管理制度。完善分类分级标准体系，建立动态调整机制，并将通用人工智能模型纳入分类分级体系，针对其涌现能力与通用能力等特殊风险，通过信息披露、监管沙箱与真实测试等治理工具建立容错机制[12]；再次，推动软硬法协同立法，在推进硬法建设的同时，大力支持行业标准、技术规范、伦理准则等软法规范的制定与完善，通过立法授权、行政认可、标准引用等方式赋予软法规范相应的法律效力，建立硬法与软法的衔接转化机制。2026 年 5 月，《国务院 2026 年度立法工作计划》首次明确提出“加快推进人工智能健康发展综合性立法”，也表明我国将“多元共治”的综合性立法列上议程。[13]

4.2 健全多元主体协同的运行机制

一是建立跨部门协同监管机构，生成式人工智能治理涉及多个部门，应在国家层面建立跨部门的人工智能治理协调机制，明确牵头部门与参与部门的职责分工，建立常态化的跨部门信息共享与执法联动机制。二是完善行业自律与社会监督机制，大力培育和发展人工智能行业组织，支持其在标准制定、行业自律、纠纷调解等方面发挥积极作用，政府通过授权委托、购买服务等方式将部分治理职能转移给行业组织；同时健全社会监督机制，畅通公众举报投诉渠道，完善舆论监督机制，发挥媒体、社会组织与公众的监督作用。[14]三是构建风险信息共享与沟通平台，建立政府部门之间的监管信息共享机制、政府与企业之间的风险信息沟通机制，以及面向社会的公共信息服务平台，打破信息壁垒。

4.3 以技术赋能协同治理

一方面要运用 AI 技术提升监管智能化水平，大力发展智慧监管技术，构建基于人工智能的风险监测、识别、预警与处置系统，实现对生成式人工智能内容的自动化检测与风险评估，提升监管的精准性与时效性，同时推动监管技术工具的共享与开放，为中小企业和行业组织提供必要的技术支持。另一方面要建立统一的风险监测与预警系统，依托国家网络安全监管基础设施，建立覆盖全国的生成式人工智能风险监测与预警系统，实现对主要平台与重点领域的常态化风险监测。

5 结论

从单向监管向协同共治的范式转换，是生成式人工智能风险治理的必然趋势，也是应对技术复杂性与不确定性的理性选择。协同共治范式以治理理论、反身法理论与动态治理理论为基础，强调主体多元化、规范协同化、机制敏捷化与责任共担化，通过多元主体的协同行动实现治理效能的整体提升。构建我国生成式人工智能风险协同共治体系，需要明确多元主体的角色定位，建立风险沟通、分类分级、软硬法衔接等核心运行机制，并从制度、机制与技术三个层面系统推进。随着技术发展与治理经验积累，协同共治的制度体系也将持续优化，最终实现技术创新与风险治理的良性互动，推动我国人工智能产业在安全可控前提下实现高质量发展。

参考文献

[1]国家互联网信息办公室等.生成式人工智能服务管理暂行办法[Z].(2023-07-13)[2026-11-11].https://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm.

[2] 姜沣格, 于海防. 生成式人工智能的法律规制——从技术与人机关系出发[J]. 求是学刊, 2025, 52(2): 123-137.

[3]施恣旻,王小雅.国枫观察丨OpenClaw 的安全边界与合规风险[OL].(2026-03-02)[2026-11-11].https://www.grandwaylaw.com/guofengshijiao/5640.html.

[4] 毕文轩. 生成式人工智能的风险规制困境及其化解：以 ChatGPT 的规制为视角[J]. 比较法研究, 2023, (3): 155-172.

[5] 陈亮, 张翔. 欧盟生成式人工智能立法实践及镜鉴[J]. 法治研究, 2024, (6): 105-118.

[6] 季卫东, 赵泽睿. 法律如何实现动态治理——人工智能不确定性的立法应对[J]. 学术月刊, 2025, 57(3): 98-109.

[7] 舒国滢. 人工智能时代的规范延迟：问题与应对——基于法哲学视角的考察[J]. 中外法学, 2026, 38(4): 845-862.

[8] 王一臻, 何伏刚. 生成式人工智能风险治理与模式完善研究[J]. 科学决策, 2024, (11): 220-237.

[9] 国家互联网信息办公室等.人工智能生成合成内容标识办法[Z].(2025-03-14)[2026-11-11].https://www.cac.gov.cn/2025-03/14/c_1743654684782215.htm.

[10] 徐继敏. 生成式人工智能治理原则与法律策略[J]. 理论与改革, 2023, (5): 72-83.

[11] 张欣. 人工智能治理的全球变革与中国路径[J]. 华东政法大学学报, 2025, 28(1): 18-32.

[12] 林洹民. 通用人工智能模型的治理策略与规则构建[J]. 当代法学, 2026, 40(5): 18-32.

[13] 国务院办公厅.国务院办公厅关于印发《国务院 2026 年度立法工作计划》的通知:国办发〔2026〕14 号[Z].(2026-05-11)[2026-11-11].https://www.gov.cn/zhengce/content/202605/content_7068345.htm.

[14] 郭朝先. 加强人工智能监管与治理[J]. 北京工业大学学报(社会科学版), 2025, 25(3): 19-25.

[1]**第 1 作者简介：**梁静（1998-），女，硕士，研究方向：人工智能法学。E-mail: 20240010@ylu.edu.cn。

[*]**通讯作者简介：**周伟萌（1982-），男，博士，教授，研究方向：人工智能法学。E-mail: 20120003@ylu.edu.cn。