

生成式人工智能行政监管的科林格里奇困境及纾解路径

刘雪凤, 丁 翩

中国矿业大学公共管理学院, 江苏 徐州

收稿日期: 2026年9月10日; 录用日期: 2026年9月28日; 发布日期: 2026年10月30日

摘 要

生成式人工智能的指数级迭代和建构性力量对既有行政监管范式形成冲击。其中, 技术迭代与行政监管回应的时序张力, 使我国行政监管面临科林格里奇困境。本文引入技科学视角重新阐释的科林格里奇困境理论, 构建“困境-纾解”分析框架。从时间、知识与权力三维度, 分别揭示技术迭代与监管滞后、技术黑箱与监管局限、技术跨域与监管碎片的困境表征, 以及三重困境的反向联动和放大效应。进而, 从技术、制度和效能维度剖析困境的发生逻辑, 阐明三因素的交织固化机理。在审视欧盟严苛监管和美国宽松监管实践得失的基础上, 融合适应治理理论和协同治理理论, 提出动态监管、适配监管与协同监管互嵌的纾解路径, 以在创新和安全的动态平衡中走出我国行政监管的科林格里奇困境。

关键词

生成式人工智能, 行政监管, 科林格里奇困境, 纾解路径

Research on the Collingridge Dilemma and Relief Paths of the Administrative Supervision of Generative Artificial Intelligence

Xuefeng Liu, Pian Ding

School of Public Policy & Management, China University of Mining and Technology, Xuzhou Jiangsu

Received: September 10, 2026; accepted: September 28, 2026; published: October 30, 2026

Abstract

The exponential iteration and constitutive power of generative artificial intelligence have disrupted the established paradigm of administrative supervision. The temporal tension between technological iteration and regulatory response has placed China's administrative supervision in the Collingridge dilemma. Drawing on the technoscience perspective to reinterpret the Collingridge dilemma, this paper constructs an analytical framework centered on "dilemma identification and relief". From the time, knowledge, and power dimensions, it identifies the manifestations of the dilemma: the misalignment between technological iteration and regulatory lag, the opacity of technological black boxes and regulatory capacity constraints, cross-domain technology and fragmented supervision, as well as the reverse linkage and amplifying effect among the three dilemmas. It further analyzes the occurrence logic of the dilemma from the technological, institutional and effectiveness dimensions, and clarifies the intertwined and solidified mechanism of these three factors. On the basis of examining the strengths and weaknesses of strict supervision in the EU and flexible supervision in the US, this paper integrates adaptive governance and collaborative governance theories, and proposes an integrated relief path featuring dynamic supervision, adaptive supervision and collaborative supervision. The aim is to resolve the Collingridge dilemma in China's administrative supervision and achieve a dynamic balance between technological innovation and security governance.

Keywords

Generative Artificial Intelligence, Administrative Supervision, Collingridge Dilemma, Relief Paths

Copyright © 2026 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 问题提出

2022 年以来, 以 Chat GPT、Deep Seek 为代表的生成式人工智能(Generative Artificial Intelligence, GAI)呈指数迭代和建构嵌入态势, 在金融、教育、公共管理等领域多面扩散[1]。面对技术变革带来的监管压力, 我国行政监管展开规制与调试。《生成式人工智能服务管理暂行办法》¹ (后文简称《暂行办法》)的施行, 标志生成式人工智能被正式纳入行政监管的范畴。然而, 我国尚未形成针对生成式人工智能的行政监管范式, 监管实践面临结构矛盾: 在技术初步应用期, 风险未充分显现, 监管部门难以精准预判并制定适效规则, 陷入“早期易控不知害”的认知困境; 在技术深度融合期, 风险规模爆发, 鉴于技术和社会已深度互嵌, 管控面临成本代价和技术依赖难题, 陷入“晚期知害不易控”的执行困境。上述时序悖论, 是技术哲学中的科林格里奇困境在行政监管领域的具象投射。

我国学者已对生成式人工智能行政监管的科林格里奇困境进行初步探索, 但成果集中于技术本身的控制困境分析或整体治理路径的探究, 缺少行政监管立场的针对研究。其不足主要体现在三方面: 其一, 现有研究多聚焦以生成式人工智能为代表的数字技术的科林格里奇困境, 如在“科学-技术-社会”的存在论情境解决困境[2]、技术控制困境的伦理引入[3]、走出生成式人工智能困境的全流程数据合规管理机制[4], 没有区分技术控制困境和行政监管困境。其二, 在行政监管立场上, 已有研究多集中于生成式

¹https://www.gov.cn/zhengce/zhengceku/202307/content_6891752.htm

人工智能的治理或监管挑战, 行政监管有所提及, 缺乏深入探究, 如数字技术对既有监管模式[5]、监管规则[6]、监管权力[7]形成挑战。其三, 在行政监管困境的路径优化上, 多数研究偏向监管框架的构建[8], 或以行政监管为主的协同治理[9], 或多方利益与全球协作下的全球生成式人工智能新范式[10]。而行政监管被置于宏观的治理框架中, 始终作为应对困境的路径之一, 缺少对其自身困境的针对纾解。上述碎片的研究范式难以有效应对生成式人工智能行政监管的规制真空和规制失效的困境, 监管方案治标不治本。

因此, 本文致力于回应我国生成式人工智能行政监管的科林格里奇困境, 在安全和发展、创新与合规间寻求平衡, 构建复合型行政监管纾解路径。具体而言: 其一, 构建“困境-纾解”的分析框架。将科林格里奇困境由技术哲学领域迁移至行政监管场域, 形成由问题诊断至对策构建的逻辑闭环。其二, 揭示我国生成式人工智能行政监管困境的现实表征与发生逻辑。基于技科学视角对科林格里奇困境理论的重新解释, 综合时间、知识和权力三个维度, 构建系统的困境识别和分析体系。其三, 探索我国生成式人工智能行政监管困境的纾解路径。在审视欧美监管实践得失的基础上, 融合适应治理理论和协同治理理论, 构建动态、适配与协同的纾解路径, 回应困境破解之需。

2. 生成式人工智能行政监管困境的分析框架构建

2.1. 科林格里奇困境的理论维度

科林格里奇困境由大卫·科林格里奇(David Collingridge)在《技术的社会控制》(The Social Control of Technology)中提出, 从时间维度揭示控制悖论: 早期因信息匮乏难以有效预防; 晚期因深度嵌入而控制困难。该理论建立在传统实证主义科技观基础上, 预设技术与社会分离, 核心观点归纳为: 时间维度表现技术迭代的线性, 知识维度强调信息的完备, 以及权力维度假定控制主体外在于技术系统[11]。科林格里奇认为, 要加强预测并熟知控制方式, 但未指明如何控制。而生成式人工智能与社会系统相嵌相生, 不是纯粹的工具力量, 故本文认为, 科林格里奇困境的原初观点与生成式人工智能行政监管不完全适配。

进而, 肖雷波和柯文(2012)引入技科学视角(发展于 20 世纪 80 年代的 STS 理论)对科林格里奇困境进行重新阐释[12], 构成本文的理论基础, 见表 1。具体内容如下: 首先, 时间维度呈现技术迭代的非线性与即时性。从非线性看, 技术与社会不是单向的时间序列, 二者是非线性的互动关系。从即时性看, 技术认知要由静态转入实践动态层面, 通过持续监测和社会学习, 及时修正错误。其次, 知识维度强调“无知状态下的决策”。事实知识和价值知识相互缠绕, 无知等价值知识是反思学习的重要构成, 需边修正边实践, 完备知识不能实现。最后, 权力维度表现为行动者的民主协商和利益博弈。技术是非人类行动者和社会行动者共同塑造的产物, 其控制牵涉开发者、企业、政府、公众、国际社会等多方主体, 需加强协同治理。

Table 1. Collingridge dilemma and its interpretation from the technoscience perspective
表 1. 科林格里奇困境与技科学视角解释的科林格里奇困境

维度	科林格里奇困境	技科学视角解释的科林格里奇困境
时间维度	技术迭代的线性	技术迭代的非线性和即时性
知识维度	知识的完备性	无知状态下的决策
权力维度	外在于技术系统的行动者	多元行动者的协商与博弈

2.2. “困境-纾解”的分析框架

基于上述分析, 本文构建生成式人工智能行政监管的“困境-纾解”分析框架, 如图 1 所示。该框

架以科林格里奇困境分析视角, 融合适应治理理论和协同治理理论。其中, 适应治理理论的动态调整和学习迭代理念为纾解时间滞后和知识局限困境提供支撑; 协同治理理论的多元力量整合理念为纾解权力碎片困境提供指引。

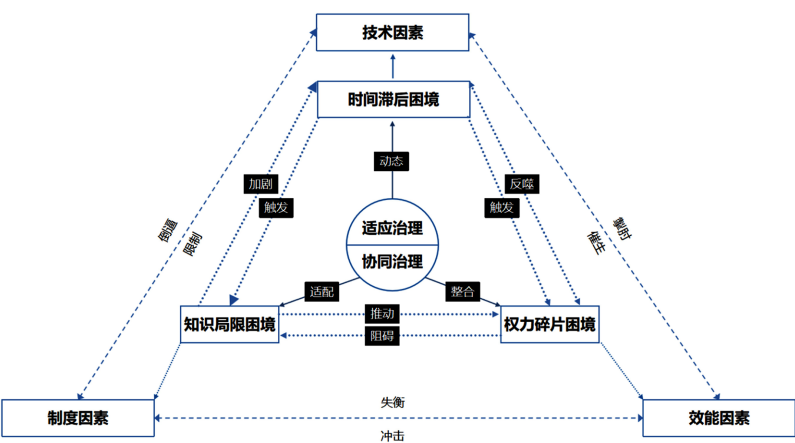


Figure 1. Analytical framework
图 1. 分析框架

2.2.1. 生成式人工智能行政监管困境的诊断逻辑

在困境诊断层面, 首先, 时间维度体现行政监管的滞后困境。受迭代差速影响, 生成式人工智能进化与行政监管回应间存在时间差, 即技术发展在前、风险显现在后。而行政监管需严格遵循程序, 其制定和执行均需制度和权力的保障。其次, 知识维度体现行政监管的局限困境。技术黑箱使监管主体难以获取关键技术信息, 无法预判潜在风险, 甚至可能错误监管。认知不足影响行政监管的制定与执行, 导致监管工具与风险样态的失配。其三, 权力维度体现行政监管的碎片困境。技术跨域特性影响监管主体的权力协调: 企业掌握关键技术信息, 政府相对处于信息被动地位。同时, 科层制下的纵向分层和横向分工致使行政监管的权责定位和资源配置相对模糊, 形成监管分化。

2.2.2. 生成式人工智能行政监管困境的纾解逻辑

在困境纾解层面, 首先, 回应时间维度的滞后困境, 构建动态监管路径。运用适应性理论中的动态调整理念, 摒弃相对静态的事后监管, 构建基于风险分类的全周期监管路径, 实现监管与技术的同频发展。其次, 回应知识维度的局限困境, 构建适配监管路径。运用适应性理论中的学习迭代理念, 建立基于法律合规的适配性监管路径, 在坚守法律硬底线的基础上加强软法更新, 完善容错和监督机制, 打造灵活的监管工具。最后, 回应权力维度的碎片困境, 构建协同监管路径。运用协同治理理论中的多元整合理念, 建立基于政府统筹的协同性监管路径, 调整政府角色定位, 推动政府领导下的多元共治和跨境协作。

因此, 文章基于科林格里奇困境理论, 构建生成式人工智能行政监管“困境 - 纾解”分析框架, 理清技术对行政监管的挑战, 实现生成式人工智能行政监管的动态、适配与协同。

3. 生成式人工智能行政监管困境的现实表征

生成式人工智能行政监管困境的本质, 是技术外溢效应与行政监管范式之间的张力。基于生成式人工智能的迭代性、黑箱性和跨域性特征, 本文将行政监管困境总结为: 时间维度下的技术迭代和监管滞后困境; 知识维度下的技术黑箱和监管局限困境; 以及, 权力维度下的技术跨域和监管碎片困境, 如图 2 所示。

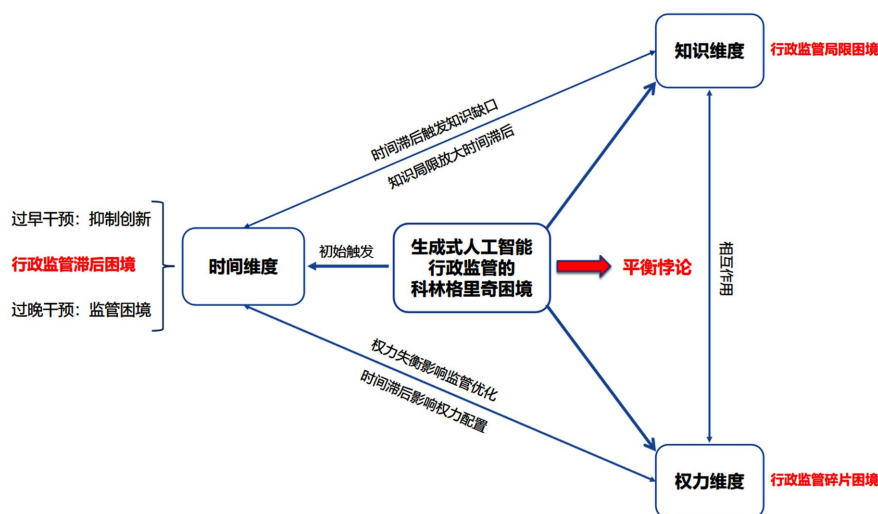


Figure 2. Collingridge dilemma in administrative supervision of generative AI
图 2. 生成式人工智能行政监管科林格里奇困境的解析

3.1. 时间维度：技术迭代和行政监管的滞后困境

生成式人工智能的动态更新导致行政监管处于滞后追赶状态，面临“看不懂”和“追不上”挑战。

首先，行政监管的稳定趋向难以与生成式人工智能的快速更新实现同频，形成技术先演化、监管后回应的初步困境。在技术初步应用期，监管脱节制约预防能力。生成式人工智能的生命周期循环加速，紧缩行政监管介入的时间和空间。而我国行政监管要遵循法定程序，兼顾程序公平和需求匹配。例如，《人工智能生成合成内容标识办法》² (后文简称《内容标识办法》)从意见稿发布到正式施行耗时 1 年。技术月度更新和监管年度调试的时间失衡，使行政监管难以实现风险预防。

其次，在技术深度融合期，规制被动导致行政监管失效。生成式人工智能融入社会系统后，技术依赖与高昂代价会纵深行政监管困境。原因在于，科层制的职能分割与生成式人工智能的跨域渗透具有矛盾。科层制的路径依赖会强化行政监管问题，如部门壁垒增加协调成本，监管虚化影响监管效果[13]。而生成式人工智能的问题追责和因果判断相对复杂，监管机制的不健全和监管措施的滞后形成增负效应，导致行政监管失灵。

3.2. 知识维度：技术黑箱和行政监管的局限困境

植根于生成式人工智能不确定性形成的黑箱效应，与监管认知、监管规则形成失衡，构成行政监管的局限困境。

首先，技术黑箱构建知识壁垒。这种壁垒表现为两点。其一，技术黑箱形成监管知识盲区。生成式人工智能算法将偏见转化为代码，歧视在技术掩盖下实现合法化。如我国《暂行办法》虽要求服务提供者说明算法原理并提供训练数据，但企业技术报告却缺少参数、风险等关键信息，是形式披露而非完全说明。其二，技术伪造造成监管场景缺口。生成式人工智能的深度升级和监管工具的线性发展形成博弈，在制度因素的作用下，技术的不确定性转化为社会风险。虽然《互联网信息服务深度合成管理规定》³加强深度伪造技术的管理，并未实现监管场景的全覆盖，依靠关键词过滤、人工检查的传统监管方式无法适配技术的跨模态特征。

²https://www.cac.gov.cn/2025-03/14/c_1743654684782215.htm

³https://www.gov.cn/zhengce/zhengceku/2022-12/12/content_5731431.htm

其次, 监管不足扩大知识鸿沟。上述挑战表现为两点。其一, 监管条款模糊引发执行偏差。我国生成式人工智能行政监管尚未成熟, 侧重于原则解释和整体规定, 缺少适配技术的具体标准。比如, 生成式人工智能会引发风险相融和预测困难的挑战, 风险分类的监管策略存在匹配难题, 既可能因为设置高合规义务阻碍行业发展, 也可能由于监管不举导致风险蔓延。其二, 监管工具失配弱化监管效果。以算法解释为例, 尝试打开算法黑箱的监管方式对生成式人工智能不适用。原因是, 一旦技术解释速度慢于技术更新速度, 算法解释则无法满足监管需求[14]。

3.3. 权力维度: 技术跨域和行政监管的碎片困境

我国生成式人工智能行政监管延续传统监管逻辑, 初步形成由中央至地方的横纵式监管结构, 与生成式人工智能跨域性形成冲突, 导致行政监管的碎片问题。

首先, 监管主体多元引发行政监管冲突。《暂行办法》推动政府主导的多部门协作和多主体参与[15], 而头部企业凭借其技术优势和经济效益, 将商业利益转化为制度诉求, 反向影响制度环境, 限制行政监管有效性[7]。与头部企业形成对比, 中小企业在生成式人工智能场域中处于弱势地位, 形成强监管、弱保护的局面。此外, 行政监管机构内部的监管干预会出现预期外影响, 表现为横向的行政部门错位和纵向的优先目标冲突。

其次, 权责配置不佳影响行政监管效能。我国生成式人工智能行政监管基本形成多方共管格局, 问题在于各部门权责划分相对模糊, 监管权的过度“行政化”会形成新问题。一方面会出现监管空缺和监管过度现象[16], 另一方面会增加协调难度并产生责任推诿。此外, 科层制的纵向执行会形成监管依赖, 横向分工会催生监管缝隙, 刚性程序会降低监管灵活性, 造成协同虚化[17]。故技术服务跨域和行政监管属地的冲突使局部风险向全局风险转化, 分散的监管策略导致行政监管陷入风险政治化和政治化风险的困境。

行政监管的三重困境并非孤立存在, 而是呈现反向联动与放大效应: 时间困境初始触发, 知识和权力困境互构, 至三维困境形成闭环。困境的互动演化使行政监管始终滞后技术创新, 使行政监管最终陷入技术创新和安全监管的平衡悖论。

4. 生成式人工智能行政监管困境的发生逻辑

本文结合我国生成式人工智能行政监管的政策文件和执行落地, 从技术、制度、效能三个维度, 解析我国生成式人工智能行政监管困境形成的深层原因。

4.1. 技术因素: 技术迭代和监管能力的适配梗阻

新技术的出现在消除原有不确定性的同时, 也会创造新的不确定性[18]。原因在于, 生成式人工智能的自进化引发潜在或显性风险, 加剧行政监管滞后, 强化时间控制难题。面对技术迭代, 行政监管因能力欠缺而在风险防范中处于被动状态。同时, 现有监管工具无法实现有效监管。在技术发展初期, 行政监管的重心偏向激发其创新价值。伴随风险衍生以及监管不适, 行政监管进而引起政府的重视。目前, 我国行政监管工具主要为备案管理和安全评估。然而, 由于生成式人工智能风险隐蔽存在, 既有工具的技术预判和预警能力不足, 行政监管仍属事后应对, 行政监管能力与技术适配间存在障碍。

4.2. 制度因素: 规制供给和监管执行的双重缺陷

监管规则的执行具有路径依赖。生成式人工智能风险推动行政监管调整, 与基于风险预防的管制思维构成矛盾。传统行政监管高度依赖惯有经验, 一旦形成风险防范的惯性思维, 会陷入固定陷阱。在时滞效应的影响下, 行政监管与生成式人工智能表面“耦合”但深层冲突[19]。例如, 我国《内容标识办法》

没有清晰界定标识主体范围, 可能导致监管对象的不确定, 影响监管执行。此外, 行政监管的条块分割与生成式人工智能的去中心化相对立, 造成权责碎片和职责重叠。纵向上, 官僚制的一元监管与生成式人工智能的多元主体相抵牾[20]; 横向上, 官僚制的部门合作受限于部门利益和组织职能, 与生成式人工智能要求的整合治理形成价值对立, 易出现各自为政。故私人利益可能代替行政监管公共目标, 而对政绩形式主义的追求, 使行政监管的有效性为人们诟病。

4.3. 效能因素: 监管目标和资源配置的实践失范

如何平衡技术创新和政策监管, 实现高质量发展和高水平互动, 是生成式人工智能行政监管的突出问题。传统监管范式既没有充分的规范标准, 也缺少权威的法律解释, 监管能力需要强化。由于技术、制度和社会的互动尚未形成稳定状态, 我国生成式人工智能行政监管在安全与创新间反复摇摆, 目标在实践中面临失衡。上述问题会导致行政监管资源的配置低效, 表现为高风险场域监管不足和严格化惩戒倾向。由于生成式人工智能的风险问责未明晰, 加之我国实行统一的监管标准, 中小企业可能因合规不足而陷入发展停滞。同时, 在算法加持的共谋下, 难以把握监管的恰当时机和干预限度, 监管机构为减弱监管失灵概率, 监管工具的组合趋于严苛和保守, 监管效能可能遭受质疑[21]。

我国生成式人工智能行政监管科林格里奇困境的发生逻辑表现为技术、制度和效能三因素的交互。技术因素构成行政监管的客观问题, 制度因素放大行政监管的知识空白, 效能因素形成监管困境的实践诱因。

其中, 行政监管困境根源于三因素的交织与固化。在双向交织关系上, 首先形成技术倒逼、制度失配的初始关系, 制度的双重缺陷也会反向限制监管工具使用; 进而形成技术与制度驱动下的效能失范的强化关系, 导致行政监管面临权责虚化、资源低效等问题, 直接催生监管目标的失衡。在系统固化关系上, 三因素表现为路径依赖加强监管制度惯性、利益博弈影响行政监管导向, 效能偏差弱化行政监管能力的关联。

5. 域外生成式人工智能行政监管的实践镜鉴

生成式人工智能的负面效应是全球范围的核心风险, 欧美等国家在应对技术挑战时同样面临行政监管困境。其中, 欧盟和美国形成对比性的监管范式, 为纾解我国的行政监管困境提供实践参照。

5.1. 域外生成式人工智能行政监管的对比

5.1.1. 欧盟: 基于风险规制的严苛监管

欧盟生成式人工智能行政监管试图以事前监管应对技术迭代, 以全面监管应对技术黑箱、以集中监管应对技术跨域, 属于安全为本、兼顾效率的监管范式。

首先, 全周期监管的事前规制路径。欧盟的行政监管导向源于重视法律的传统, 试图以统一的《人工智能法》⁴ (后文简称《法案》) 来实现强预防监管[22]。其中, 欧盟将人工智能风险界定为四个等级, 并规定以生成式人工智能为代表的高风险人工智能系统的义务[23]。同时, 欧盟构建宏观统筹、中层助推、下层落地的监管架构[23], 并推行监管沙盒。其次, 穿透式监管的规则预设路径。欧盟重视立法与监管的统一, 借助《通用数据保护条例》⁵《数字服务法》⁶等法律, 涵盖欧盟内的所有市场和欧盟外的数据传送[24]。欧盟也设立专门的监管机构, 重点部署人工智能关键领域, 完善高额违规条款和侵权责任认定。最后, 集权主导的多层次监管架构。欧盟人工智能委员会发挥统领作用, 各成员国监管机构负责具体执行,

⁴<https://aisg.tongji.edu.cn/info/1005/1192.htm>

⁵<https://secnotes.cn/grc/ai/gdpr/>

⁶<http://data.europa.eu/eli/reg/2022/2065/oj>

在辅助性原则下保留回旋余地, 允许合法基础上的自我调试。

综上分析, 欧盟以全过程风险管理保障人工智能安全。然而, 刚性的前置监管与生成式人工智能的迭代具有冲突, 激化安全与创新的紧张关系, 一定程度上体现欧盟监管的创新抑制效应。

5.1.2. 美国: 基于场景应用的宽松监管

美国基于生成式人工智能涟漪影响和维持人工智能领先的需要, 行政监管以事后规制、经验迭代、共治监管为特征, 属于效率优先、兼顾安全的监管范式。

首先, 软性监管的事后规制路径。美国的行政监管呈现由行业实践到行政指引, 再至立法确认的渐进特征, 强调问题导向的场景化监管, 如 2025 年纽约州的《培育原创、促进艺术和保障娱乐行业安全法案》⁷关注生成式人工智能伪造问题, 以减少未经授权的模仿。其次, 经验迭代的动态监管路径。美国以打补丁形式推进监管, 延续普通法传统下的经验主义知识生产。从奥巴马政府时期的《为人工智能的未来做好准备》⁸等文件, 至特朗普政府时期的《赢得竞赛: 美国人工智能行动计划》⁹等文件, 再到拜登政府试图构建的“一体四翼”监管模式^[25], 都体现美国对创新激励的重视, 而未出台统一的行政监管法案。最后, 分权制衡的共治监管架构。美国形成联邦和州的分层式监管体系, 联邦没有成立专项监管机构, 各州呈现拼图式监管格局。此外, 各州甚至市政府充当人工智能监管实验者, 如加利福尼亚州在隐私法中规定算法透明要求。

综上分析, 美国更强调发挥技术价值, 适应性监管具有强操作性。然而, 重创新轻监管的实践不利于国内监管平衡和全球共生发展, 缺乏法治化的底线约束使生成式人工智能行政监管之路道阻且长。

5.2. 域外生成式人工智能行政监管的启示

基于我国安全与发展并重的行政监管导向, 立足生成式人工智能行政监管语境, 从监管介入时序、监管规制措施、权力配置方式三个角度, 总结适合我国实际的行政监管启示。

5.2.1. 监管时序: 构建审慎与动态的监管体系

欧美生成式人工智能监管时序的路径表明, 事前规制和事后规制均具有局限。双方的共性是试图实现风险防控与安全保障的同步发展, 故强化技术可控是破解算法黑箱的关键。我国可以双管齐下, 既增强生成式人工智能算法的可解释性, 又借鉴美国场景监管的工具包, 开发算法监管工具。此外, 精密风险分级体系和动态调整机制的结合是实现有效监管的基石。我国可以吸收《法案》的风险管理思维, 以风险的识别评估、全周期管控、适应调整为遵循, 完善监管规则并加强动态调试, 既坚守安全底线, 又为技术创新预留合理空间。

5.2.2. 监管知识: 建立兜底和弹性的监管制度

欧盟的集中监管与美国的分散监管实践, 表明规则确定性与制度适应性平衡的重要性。而立法先行与行业自律的监管模式利弊共存, 要与国情深度结合并综合调整。面对欧盟教训, 我国行政监管要调整监管策略, 建立激励相容机制, 如调整中小企业的合规程序, 或设定合理的豁免条款, 或提供技术援助或成本补贴。面对美国的经验, 行政监管不仅要关注准入环节, 还要关注市场竞争的公平与否。因此, 我国需探索刚柔并济路径, 国家层面要构建统一的监管范式, 坚守监管防线; 同时赋予监管机构弹性空间以实现横纵监管, 应对监管的知识局限。

⁷<https://ipr.mofcom.gov.cn/article/gjxw/gbhj/bmz/mg/202408/1987261.html>

⁸<https://d.wanfangdata.com.cn/periodical/ChpNaW5lcBlcmVZGljYWxDSEkyMDIyMTIwN-hlSeHhcXl0eGJtMjAxNjEyMDAyGghhZHQ0bjZpZA%3D%3D>

⁹https://www.ciste.org.cn/gjkwj/zkgd/art/2025/art_15713f49682a49ff9ae1aadde0031fd3.html

5.2.3. 监管权力：搭建统筹与协同的共治体系

欧盟和美国的监管实践都凸显统筹和协同的重要性。欧盟人工智能办公室的统筹和各国监管机构的执行能形成相对稳定的监管格局，避免条款分割的责任推诿。美国发展人工智能理念下的分散监管，在促进创新的同时能审慎处理生成式人工智能风险，为我国政府监管提供新思路。对此，我国可以健全政府负责、社会协同、公众参与的协同治理网络，并辅以常态化的跨部门协调和沟通机制，以多方合作促进监管完善。

因此，欧美生成式人工智能监管实践表明，有效监管不是严苛与放任的单向取舍，要超越事前与事后、穿透与底线、集权与分权思维的对立，实现创新与安全的双向赋能。

6. 生成式人工智能行政监管困境的纾解路径

基于欧美行政监管实践得失，结合我国语境，融合适应治理理论和协同治理理论，本文构建兼具动态、适配和协同的复合行政监管路径，如图 3 所示。

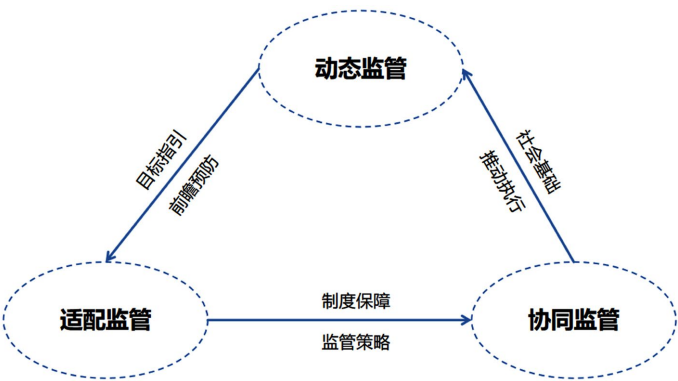


Figure 3. Pathways to alleviate the Collingridge dilemma in China's administrative supervision of generative AI
图 3. 我国生成式人工智能行政监管科林格里奇困境的纾解路径

6.1. 基于风险分类的全周期行政监管路径

动态监管路径旨在回应时间困境，从生成式人工智能特性出发，以适应治理的动态调整为理念，试图实现监管和技术的同频。

首先，建立技术成熟度与风险分类的动态目录。我国生成式人工智能官方机构可以联合智库和科研机构，制定并定期更新《生成式人工智能风险分级分类指南》，划分实验级、受限级、商用级和禁止级层级，辅以差异化监管清单，建立季度更新机制。例如，实验级应用的监管以伦理审查、软法引导和定期追踪为主；受限级应用的监管主要是准入审查、强制审计和持续监督；商用级应用关注全程监管、合规激励和违规惩罚；禁止级应用则严格禁止一切研发行为。

其次，加强生成式人工智能的全周期闭环监管。一是行政监管人员要熟知技术不同阶段的风险类型和程度，选择合适的监管时机，以不影响技术创新为基线。二是制定生成式人工智能闭环监管的措施。事前预防阶段要坚持源头治理，网信办可以牵头建立生成式人工智能风险数据库，并根据合规清单来开展备案与评估。事中管控阶段要搭建行政监管监测平台，设置分级分类的风险预警指标及对应系数，对不同风险等级的企业实施差异监管。事后优化阶段要制定总结举措并加强沟通互动。三是设立生成式人工智能行政监管的观察期。监管机构要全面收集生成式人工智能的风险信息，为监管介入提供支撑[26]。

6.2. 基于法律合规的适配性行政监管路径

适配监管路径旨在回应知识困境, 以适应治理框架内的学习迭代为理念, 提升行政监管的学习和调适能力。

首先, 完善刚柔并济的行政监管制度。一是刚性监管制度层面, 明确监管原则、规范主体和监管内容。在监管原则上, 监管机构坚持创新和安全结合的逻辑。在规范主体上, 搭建贯穿技术研发主体、服务供给主体和用户的全链条风险问责制度。在监管内容上, 关注分散的监管内容, 补充模糊或过时的内容。二是柔性监管制度层面, 弥补硬性监管的滞后弊端, 表现为国家立法和社会规范。前者即国家法律文本和规范文件的柔法, 如美国《关于安全、可靠和可信人工智能开发与使用的行政命令》¹⁰; 后者即行业规定、团体规定或生成式人工智能在开发应用过程的细则等, 如我国学者公布的《人工智能示范法 2.0 (专家建议稿)》¹¹。

其次, 健全动态适配的行政监管机制。一要构建容错机制, 激励生成式人工智能的创新。企业等监管对象的容错次数只有一次, 适用条件为主动报告合规偏差、未造成严重后果, 且能主动整改的企业。同时, 审批流程可以参照已执行的监管流程, 审批时间控制在适当期限内。对于政府监管机构, 允许一定范围内的错误, 并辅以容错复盘和人员培训。二要完善监督机制, 保证生成式人工智能规范监管。为避免监管机械执行的失效, 政府监管机构拥有适量裁量权, 而权力可能被滥用。因此, 在监督监管对象之余, 亦应构建对监管机构的监督机制。政府监管活动要与生成式人工智能行政监管平台互通, 一旦监测到不合规的监管行为, 系统会触发预警并传达上级监管机构。同时, 监管需要司法审查的支持, 对于不当的行政监管, 司法机关要评估监管行为的正当性和裁量的程序性。

最后, 打造灵活高效的技术监管工具。一要优化多元监测体系。监管政策立废与否, 要接受实践检验并优化调整。我国政府监管机构可以引入社会化评估模式, 建立监管政策的快速通道, 对不适应技术发展或负面效应显著的条款, 及时调试或废止; 对符合监管需求的条款, 加以完善。同时, 提升数字政府水平, 加大监管科技投入。监管机构要加强自身能力建设, 建立大数据辅助决策与监管的机制, 也要培养懂法律、熟技术的复合型人才。二要推进监管工具的试点工作。我国行政监管可以采纳监管沙盒的测试机制, 将累积信息转化为监管策略, 从而形成监管的政策演化和制度调试。在实行监管沙盒之余, 也要综合筛选试点区域并加强资源供给, 为行政监管的试点工作提供有力支撑。

6.3. 基于政府统筹的协同性行政监管路径

协同监管路径旨在回应权力困境, 以协同治理的力量整合为理念, 实现互斥向共治的转变。

首先, 构建政府主导的协同监管机制。政府要从单一监管向“元监管”转变。为应对我国行政监管的权责分散和资源低效问题, 政府内部要强化顶层设计, 完善生成式人工智能监管准则^[27], 并设立跨区域、跨部门的监管协调制度, 定期开展人工智能监管活动。同时, 构建政府 - 行业 - 公众的对话机制, 尝试建立常态的主体对话平台, 借助听证会、研讨会等, 完善沟通渠道。此外, 生成式人工智能监管离不开企业和行业组织的参与, 上述主体要制定行业自律准则和实践指南。

其次, 推动监管规则对接和跨境协作。生成式人工智能行政监管尚未形成统一的全球监管范式, 需建立合作关系。一要搭建跨国监管规则平台, 建立多边监管联盟。在尊重各国法律主权的基础上, 尝试建立全球性或区域性的人工智能监管协调平台, 实现各国监管标准的对接。在此基础上, 各国监管者要加强动态协调, 如建立全球生成式人工智能行政监管联盟, 定期评估人工智能风险并讨论监管方案, 推

¹⁰<https://bidenwhitehouse.archives.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/>

¹¹<https://aisg.tongji.edu.cn/info/1005/1211.htm>

动有效经验上升为约束规则。二是积极整合多方监管力量, 构建多方参与的治理网络。生成式人工智能的国际监管要鼓励多主体的民间合作, 以弥补政府监管的不足。为此, 可以建立行业主导的国际人工智能监管联盟, 加强跨国对话, 以各领域的监管合规案例为切口, 总结有效经验, 共同制定准则。

因此, 动态监管、适配监管、协同监管的纾解路径互嵌, 共同回应生成式人工智能行政监管的科林格里奇困境。动态监管路径着力解决行政监管的时间选择难题, 适配监管路径旨在提升行政监管的适应敏捷性, 协同监管以整合力量应对监管碎片问题, 最终提升生成式人工智能行政监管效能。由此可见, 三维监管环环相扣, 共同聚焦于在不确定的环境中构建既坚守安全底线、又激发技术创新效能的中国特色的生成式人工智能行政监管路径。

7. 结论

本文引入技科学视角重新阐释的科林格里奇困境理论, 构建“困境-纾解”分析框架, 揭示生成式人工智能行政监管在时间、知识、权力三维度的困境表征, 阐明技术、制度、效能三因素交织固化的发生逻辑, 并提出动态、适配与协同互嵌的纾解路径。因此, 本文的核心价值在于: 将科林格里奇困境从技术哲学迁移至行政监管场域, 拓展生成式人工智能行政监管的研究视野; 构建从问题诊断至路径纾解的完整框架; 并提出具有中国语境特色的复合监管路径, 为技术创新与安全监管的平衡提供基础的理论工具与实践参照。未来, 可以进一步探究生成式人工智能具体应用场景中的行政监管困境, 借助数据建模、案例研究等方法, 验证文章纾解路径的适效性。

基金项目

教育部人文社会科学研究规划基金“生成式人工智能应用中知识产权安全风险的生成机理与治理路径研究”(25YJA630051)。

参考文献

- [1] 胡登峰, 吴昊, 王佳怡. 人工智能时代政府适应性治理研究——基于“技术-制度-价值”三维分析框架[J]. 学术月刊, 2025, 57(6): 83-94.
- [2] 陈凡, 贾璐萌. 技术控制困境的伦理分析——解决科林格里奇困境的伦理进路[J]. 大连理工大学学报(社会科学版), 2016, 37(1): 77-82.
- [3] 文成伟, 汪姿君. 预知性技术伦理消解 AI 科林格里奇困境的路径分析[J]. 自然辩证法通讯, 2021, 43(4): 9-15.
- [4] 张春春, 孙瑞英. 如何走出 AIGC 的“科林格里奇困境”: 全流程动态数据合规治理[J]. 图书情报知识, 2024, 41(2): 39-49+66.
- [5] 杨松, 张永亮. 金融科技监管的路径转换与中国选择[J]. 法学, 2017(8): 3-14.
- [6] 刘权. 数字经济视域下包容审慎监管的法治逻辑[J]. 法学研究, 2022, 44(4): 37-51.
- [7] 余南平, 栾心蔚. 论人工智能监管——国家-市场关系视角下的人工智能技术权力[J]. 世界经济与政治, 2025(6): 31-59+154-155.
- [8] 韩松言, 黄静. 何以破解生成式人工智能监管的科林格里奇困境?——一个包容审慎框架的构建[J]. 西华大学学报(哲学社会科学版), 2025, 44(6): 55-65.
- [9] 沈芳君. 生成式人工智能的风险与治理——兼论如何打破“科林格里奇困境”[J]. 浙江大学学报(人文社会科学版), 2024, 54(6): 73-91.
- [10] 方兴东, 宋珂扬, 钟祥铭. Deep Seek 与科林格里奇困境: 人工智能失控的风险机制及治理可能[J]. 新闻与写作, 2025(5): 51-59.
- [11] Collingridge, D. (1980) *The Social Control of Technology*. Open University Press, 11+12+16-17.
- [12] 肖雷波, 柯文. 技术评估中的科林格里奇困境问题[J]. 科学学研究, 2012, 30(12): 1789-1794.
- [13] 周雪光. 权威体制与有效治理: 当代中国国家治理的制度逻辑[J]. 开放时代, 2011(10): 67-85.

- [14] 周翔. 算法可解释性: 一个技术概念的规范研究价值[J]. 比较法研究, 2023(3): 188-200.
- [15] 张树华, 张夏添. 生成式人工智能的智慧监管: 理论、模式与选择[J]. 中国社会科学院大学学报, 2025, 45(5): 26-50+174-175+181.
- [16] 张欣. 生成式人工智能的算法治理挑战与治理型监管[J]. 现代法学, 2023, 45(3): 108-123.
- [17] 彭勃, 林铭海. 人工智能应对国家安全风险: 科层化治理的技术重塑之路[J]. 上海交通大学学报(哲学社会科学版), 2025, 33(10): 15-32.
- [18] 斯万·欧维·汉森, 刘北成. 知识社会中的不确定性[J]. 国际社会科学杂志(中文版), 2003, 20(1): 37-45.
- [19] 张彦华, 徐帆. 人工智能影响政治安全的逻辑及其风险治理——基于传播政治经济学的分析[J]. 社会科学战线, 2022(12): 196-205.
- [20] 谭九生, 杨建武. 人工智能技术的伦理风险及其协同治理[J]. 中国行政管理, 2019(10): 44-50.
- [21] 崔小燕, 张长立. 生成式人工智能诱发的风险样态、生成逻辑与智慧治理策略——基于传播政治经济学的分析[J]. 宁夏社会科学, 2023(4): 188-198.
- [22] 陈凤仙, 连雨璐, 王娜. 欧美人工智能监管模式及政策启示[J]. 中国行政管理, 2024(1): 77-88.
- [23] 何野. 生成式人工智能安全监管制度优化: 欧美经验及其启示[J]. 中国科技论坛, 2025(10): 180-188.
- [24] 李佳轩, 常帅峰, 袁勤俭. 中外生成式人工智能监管政策比较研究[J]. 情报理论与实践, 2025, 48(6): 37-46.
- [25] 邢亚杰, 戚凯. 论当前美国政府的人工智能监管政策[J]. 国际观察, 2024(4): 31-57.
- [26] 卢超. 包容审慎监管的行政法理与中国实践[J]. 中外法学, 2024, 36(1): 143-160.
- [27] 任梅, 代栋栋. 生成式人工智能的回应性监管架构及其实现路径[J]. 中国行政管理, 2025, 41(2): 39-50.