doi. 10. 3724/1005-0566. 20260218

适时与适度：
人工智能监管的理论框架构建及政策路径选择

魏　巍[1],冯　翼[2],刘　蕾[3]

(1. 国家发展和改革委员会市场与价格研究所,北京　100038)

(2. 成都市经济发展研究院/国宏访问学者,四川　成都　610032)

(3. 北京工商大学商学院,北京　100048)

摘　要:针对人工智能技术非线性演化、跨域渗透引发的“监管时机困境”与“强度失准风险”等治理赤字,传统静态、线性的监管范式已难以适配,深陷科林格里奇悖论所揭示的“控制困境”。本文构建“适时性与适度性”耦合的动态治理框架。基于颠覆性技术监管轨迹梳理与欧盟、美国、英国、中国比较,发现“信息不足、控制受限”仍为关键约束,单纯预防或放任均难适配。进而提出以社会渗透率和场景敏感性触发三阶段滚动介入,并以比例原则形成四级风险工具箱,配套评测审计、责任链条及国际程序互认,实现创新与安全的动态均衡。

关键词:人工智能治理;适时适度监管;比例原则;社会渗透率;科林格里奇悖论

中图分类号:　　**文献标识码**:A　　**文章编号**:1005－0566(2026)02－0210－15

Right timing, right intensity:
Constructing a theoretical framework for AI regulation and policy pathways

WEI Wei[1], FENG Yi[2], LIU Lei[3]

(1. *Institute of Market and Price*, *National Development and Reform Commission*, *Beijing* 100038, *China*; 2. *Chengdu Economic Development Academy / Visiting Scholar*, *China Academy of Macroeconomic Research* (*CAMR*), *Chengdu* 610032, *China*; 3. *School of Business*, *Beijing Technology and Business University*, *Beijing* 100048, *China*)

Abstract: The rapid, nonlinear evolution and cross-domain penetration of artificial intelligence (AI) have precipitated significant governance deficits, manifesting as “timing dilemmas” for regulatory intervention and “miscalibrated intensity risks”. These challenges expose the limitations of traditional static and linear regulatory paradigms, trapping them within the “control dilemma” articulated by the Collingridge dilemma. This paper constructs a dynamic governance framework predicated on the coupling of temporality and proportionality. Through a historical analysis of disruptive technology regulation and a comparative study of the European Union, the United States, the United Kingdom, and China, the research identifies persistent key constraints: insufficient information and limited control, rendering purely precautionary or laissez-faire approaches inadequate. Consequently, the paper proposes a three-stage rolling intervention mechanism triggered by social penetration rate and contextual sensitivity, alongside a four-tier risk-calibrated toolkit grounded in the principle of proportionality. This framework is supported by complementary measures such as evaluation

基金项目:中国宏观经济研究院(国家发展和改革委员会宏观经济研究院)2025 年基本经费研究课题;中国博士后科学基金第 77 批面上项目(2025M773676)。

作者简介:魏巍(1990—),男,青海西宁人,国家发展和改革委员会市场与价格研究所助理研究员,北京工商大学博士研究生,研究方向为数据要素、人工智能及宏观经济等。通信作者:刘蕾。

and auditing systems, clear accountability chains, and international procedural mutual recognition, aiming to achieve a dynamic equilibrium between fostering innovation and ensuring safety.

Key words: AI governance; timely and proportionate regulation; principle of proportionality; social penetration rate; Collingridge dilemma

过去10余年间,人工智能(AI)技术经历了从深度学习到大语言模型(LLMs)及生成式人工智能(GAI)的跨越式演进。2012年,AlexNet的出现引领深度学习范式,推动了学术界与产业界的发展;2016年,AlphaGo在围棋人机大战中击败世界冠军李世石;2017年,Google提出的Transformer架构根本性变革了自然语言处理领域;2020年,GPT-3和扩散模型取得重大进展;2022年11月,ChatGPT的发布全面掀起了GAI热潮。目前,AI能力呈指数级迭代,应用场景从语音识别、医学影像扩展至自动驾驶、智能制造等众多行业。然而,AI在注入经济社会发展动力的同时,也凸显了算法偏见、数据滥用和自动化决策不透明等风险,系统的安全性、可控性和可解释性已成为社会焦点。在发展动能与安全风险交织的背景下,实现高质量发展与高水平安全的良性互动,已成为智能时代公共政策与技术监管的核心命题。

为此,全球主要经济体纷纷出台AI监管政策。欧盟《人工智能法案》(EUAIAct)、美国《AI权利法案》、中国《算法推荐管理规定》等法规的密集发布,标志着"治理先行"已成为国际共识。但正如技术创新理论学者科林格里奇所洞察到的,新兴技术往往面临着一个根本性的治理悖论,即在技术发展初期,虽然控制相对容易,但人类对其潜在影响缺乏充分认知;而当技术的深层影响逐渐显现并被大众所接受理解时,对其实施有效控制却变得异常困难。对于AI这样空前未有之颠覆性技术而言,围绕公共权力何时介入、以何种强度与边界介入,成为各国政府面临的重大理论和实践问题。为避免概念漂移,本文所称"监管"限于公共权力主导的规则设定与执行问责链条;"治理"则是包含标准、评测、行业自律与公众参与在内的更宽协作结构。①

从理论层面看,AI监管的核心并非"政府干预与市场自由"的二元对立,而是制度性公共干预如何实现动态适配。传统监管理论侧重正当性(何以监管)、工具选择(如何监管)以及绩效评估(监管效果),其预设前提是技术环境稳定且边界清晰[1]。然而,AI技术的非线性演化、跨域渗透及高度不确定性,使传统模式面临三重挑战[2]。一是时机选择的动态性。AI从研发到应用的周期急剧压缩,传统"反应式"监管面临滞后则造成不可逆影响、过早介入则因信息不足导致误判的两难困境。二是监管强度的比例性。面对AI技术应用场景与风险等级的高度异质性,传统"一刀切"或分业监管模式难以适应多维复杂性,无法实现手段与风险的精准匹配。三是监管能力的认知局限。AI"黑箱"特性及严重的信息不对称导致监管决策面临"知识赤字"。由此要求监管不能仅依靠政府单方面的能力建设,而需构建包括技术开发者、行业组织、学术机构、公民社会在内的知识共享与协同监管网络,让市场主体和社会力量成为政府可用、能用、好用的"治理资源",进而作为监管有效性的必要条件被纳入多元协同治理框架。

从实践层面看,当前全球AI监管尚未形成统一规制思路,"治理赤字"在地缘政治博弈、行业自律失灵与法律适配性危机交织下持续加剧。一是规则碎片化导致监管套利,主要经济体基于差异化战略考量形成割裂的规制版图,引发"逐底竞争"风险。二是软法治理乏力,依赖行业自律与伦理承诺的模式在商业压力下易沦为"伦理洗白",推动监管向强制性硬法收紧。三是AI的通用性与跨域渗透对既有著作权法、劳动法等垂直法律体

① 本文所称"监管"(regulation),是指政府及其法定监管机构通过制定并执行具有一般约束力的规则与合规义务(并以监督问责与救济机制保障其实施),对AI研发、部署与使用行为进行持续且聚焦的制度性约束;"治理"(governance)则指政府、市场与社会多元主体通过标准、评测、行业自律与公众参与等机制共同塑造技术运行秩序的更宽制度安排。下文中,凡讨论规则义务与执法问责,使用"监管";凡讨论多主体协作与能力供给,使用"治理"。

系造成系统性冲击,产生类似“破窗效应”。四是合规成本分布失衡,复杂的监管要求在科技巨头与初创企业之间形成不对称壁垒,加剧市场集中与创新受抑的“马太效应”。同时,传统监管政策制定与执行的冗长程序难以匹配 AI 的快速迭代,“监管真空”与“监管过度”现象交替出现。

鉴于此,监管介入的时机与力度直接关系到 AI 能否在可控边界内实现社会价值最大化。尽管在 AI 伦理、算法公平性和数据隐私保护等领域已有颇为丰富的研究成果,然而围绕监管时序与强度的系统性理论分析仍显匮乏。现有文献多聚焦特定场景或行业的监管实践,停留于对各国政策工具的描述性比较,未能深入揭示介入时机与监管强度之间的动态耦合关系。本文认为,AI 监管本质上是一个多维度、可迭代的动态优化过程,需基于技术成熟度与潜在风险的阶段性演进设计差异化策略。从这一认识出发,“适时”意味着把握技术发展的关键节点,选择最佳的监管介入时机;“适度”则指根据技术的社会渗透率、风险水平和社会接受度等设定合理强度,两者相互制约,共同构成 AI 监管的核心要素。基于此,本文立足当代科技治理困境,回溯技术监管历史脉络,前瞻智能时代治理图景,通过剖析技术演进规律与监管逻辑的内在关联,探寻 AI 治理的时度把握之道,以期为新兴技术治理的政策设计提供理论参照。

一、技术治理的历史逻辑与监管时序演进

二战后,随着科技活动对国家综合实力的影响显著增强,各国政府逐步以有组织、制度化的方式介入科技发展进程[3]。在此背景下,技术治理实践与理论持续演进,形成典型监管实践,并引发科林格里奇悖论等标志性命题及理论反思。

(一)颠覆性技术监管的历史轨迹分析

“颠覆性技术”可谓是基于最前沿的技术突破或创新组合带来变革式应用效果的新技术[4]。虽然近代科学在 17 世纪已经诞生,但彼时科技活动对国家经济社会发展的影响并不明显,政府主要是引导和鼓励研究主体将科研方向与国家发展所需相结合,基本未进行实质性干预[5-6]。二战期间,美国等国通过核武器、雷达等颠覆性技术实现突破,凸显科技对国家竞争力的重要性,推动各国健全相关法律法规与监管体系[6]。以核技术为例,美、俄、英、法等拥核国出台专项法规、标准,建立专责机构,并根据技术演进与国际形势持续调整国际条约与国内规制,实现严格管控[7]。

进入 21 世纪,随着新一代信息技术的发展,科技创新及其生态发生了根本性变革:参与主体多元化,创新引领从政府主导向政府、企业、高校、科研院所多维协同转变,技术迭代周期显著缩短,新技术、新产业、新业态、新模式不断涌现,颠覆性技术对经济社会各领域产生的影响更为广泛而深刻。不同国家基于差异化的发展考量,在颠覆性技术监管策略的选择上也呈分化态势,可归纳为 3 种典型路径:一是以美国、以色列等为代表的技术优先型国家,强调在全球科技竞争中抢占制高点、最大化释放技术创新活力,倾向于通过减少事前管制、简化审批程序、提供政策激励等方式[8-9],使技术在相对宽松自由的环境中快速演进,待风险显现后再进行针对性治理,其核心逻辑是“先发展、后规范”;二是以德、法等欧洲经济体为代表的安全导向型国家,更加关注技术发展可能带来的社会风险与伦理挑战,重视先行构建相对严格的事前监管框架,通过建立技术准入门槛、强化合规要求、加强风险评估等手段[10-11],确保技术发展始终在可控范围内,其治理哲学体现为“审慎发展、风险优先”;三是以英国、新加坡、韩国等为代表的试验探索型国家,试图在创新激励与风险防控之间寻求平衡点。这一路径的核心制度创新是“监管沙盒”机制的建立与扩散。英国金融行为监管局(FCA)于 2016 年率先推出“监管沙盒”开创了以限期、限域与有限豁免为核心的试验性框架[12];新加坡金融管理局(MAS)同年跟进设立金融科技监管沙盒,并于 2019 年推出简化版 Sandbox Express 以加速创新项目落地;韩国于 2019 年建立 ICT 规制沙盒;日本则在 2018 年启动内阁府“单一窗口”沙盒并于 2021 年实现常设化。上述国家通过监管陪跑与数据化评估,力求实现“适时介入”与“适度约束”的动态平衡。该模式的成功经验随后被法国能源监管、澳大利亚金融监管以及中国金融科技创新监管试点等多个场景借鉴采纳,成为全球监管创新的重要范式。

表 1　主要国家/地区监管沙盒实施情况对比

国家	启动年份	沙盒项目、覆盖领域和主管机构	申请条件(要点)	试验期限	核心特征
英国	2016	金融科技监管沙盒,由英国金融行为监管局(FCA)主管	创新性、消费者受益、测试准备就绪	一般不超过 6 个月	全球首创监管沙盒;配备专门的沙盒团队与规则工具箱
新加坡	2016	金融科技监管沙盒,由新加坡金融管理局(MAS)主管	创新性、测试边界明确、可受益于监管豁免	一般根据个案确定;快捷沙盒可试验 9 个月	在一般性监管沙盒基础上,逐步推出快捷沙盒(允许更快启动测试)和沙盒 +(创新辅助服务更为完善);强调有限豁免和明晰边界
德国	2019	现实实验室,覆盖能源转型、交通出行、数字医疗等领域,由德国联邦经济与气候保护部(BMWK)主管	创新性、现行法规受限	一般不超过 4 年	强调“试点—评估—规制学习”;推进“Reallabor 法”
韩国	2019	信息通信技术(ICT)规制沙盒,由韩国科学技术信息通信部(MIST)主管	创新性、商业模式明确、现行法规受限	一般不超过“2 + 2”年(即常规 2 年 + 延长 2 年),后续还可延长至相关立法完成	“先行后规制”;快速通道、临时许可与示范豁免并用
法国	2019	能源领域监管沙盒,由法国能源监管委员会(CRE)及相关主管当局实施	创新性、风险可控、现行法规受限	一般不超过 4 年,可延期 1 次	重视数据与网络安全;分阶段开展效果评估
澳大利亚	2020	增强监管沙盒,用于金融领域,由澳大利亚证券与投资委员会(ASIC)主管	创新性、消费者受益、测试计划完备	一般不超过 24 个月	沙盒内的许可豁免范围广;允许试验时间长,进入沙盒成本低
日本	2018	监管沙盒系统,覆盖金融科技、自动驾驶、无人机、医疗等领域,由内阁府/内阁官房“单一窗口”,与各省厅协同实施	创新性、具有社会效益、现行法规受限	按项目设定试验限期,可延期	跨部门协调实施;设有评价委员会;试验数据用于法规检讨
中国	2019	金融科技创新监管试点,由中国人民银行会同相关部门实施	创新性、服务实体经济、依法合规、普惠性、保护消费者权益	一般为 1 年	强调包容审慎;分批实施

说明:上表统一优先采用官方/权威机构口径;“期限”一栏如为“常见上限/个案确定”,表示无统一硬封顶、以批复/个案为准。澳大利亚曾于 2016 年首次推出监管沙盒,此表仅呈现其于 2020 年推出的增强监管沙盒情况。
资料来源:FCA、MAS、MSIT、日本内阁府、法国 CRE、ASIC 及中国人民银行等官方文件与权威解读。

(二)科林格里奇悖论的现实表征与理论反思

虽然各国针对科技变革下的颠覆性技术采取诸多监管举措,但科林格里奇悖论所揭示的“信息不足与控制受限”的双重约束仍具有解释力,在实践层面至少可从以下 3 个方面观察其持续影响。一是监管革新与技术发展的时间错位。颠覆性技术的诞生对监管体系带来变革课题,但监管者既受限于对新技术的知识缺乏,也受制于现有制度约束,易导致“技术—监管时滞”[2]。多国在金融科技领域的追赶式监管即为例证[13]。二是监管措施与技术特性的契合偏差。受技术发展不确定性、认知局限性、传统监管理念与方式的路径依赖等因素影响,监管者往往难以针对颠覆性技术的特点精准配置监管举措,进而导致监管真空、监管乏力或监管过度。例如,部分监管者早期对基因编辑技术的影响研判失准,放开个别技术成果问世后引发全球性伦理争议;后又以“急刹式”高强度禁令仓促应对舆论压力,始终未能弥合监管与新技术之间的适配性鸿沟[14]。三是监管效力与技术影响的效果落差。当部分技术深度嵌入社会运行机理后,其所构建的经济利益链条以及对使用者形成的锁定效应共同构成监管实施的结构性阻力,并引发监管成本与收益间的平衡难题。例如,平台经济全面渗透至信息传播、商业流通等领域后,即便是美国等发达经济体,仍在数据保护、市场反垄断等监管领域与平台巨头陷入持久拉锯。

科林格里奇提出“控制困境”,但其立场并非悲观论[15];相反,其论证旨在引出应对路径。随着技术演进与监管革新的互构实践持续深入,学界关于科林格里奇悖论的反思逐步展开。第一,科林格里奇悖论预设技术线性发展而非动态演化,遮蔽了监管介入的多重时间窗口选择。科林格里奇认为技术发展遵循“基础科学—产业应用—社会扩散”的单向线性演进路径[16],因此监管者须把握技术扩散前的某一最优时点介入。但现实中,技术发展多呈现非线性演进特点,且当前许多

技术创新随社会需求动态重构,技术改进与市场反馈动态耦合,为监管带来了更多可介入的节点选择。第二,科林格里奇悖论强调预知技术发展而非迭代技术认知,忽视了监管措施的渐进调适可能。科林格里奇推崇“全知决策”,即只有在监管介入前全面掌握技术知识、准确预测技术发展,方能作出合理决策[17]。这一判断既背离了技术的可证伪性,也偏离了人类对技术认知渐进深化的一般规律[18]。在“技术演化—知识迭代”的动态循环中,监管策略实则可跟进调适,逐步提高与技术特性的匹配性。第三,科林格里奇悖论侧重控制技术而非治理技术,缺失了多元主体共同参与的协作共赢想象。科林格里奇认为监管者之于技术是一种控制关系,与其他技术相关主体是对立角色。然而,监管者需要向技术创新者、使用者等征询技术知识;技术创新者、使用者等也需要监管者维护良好的技术发展环境。此种双向依赖关系能够促使技术监管者与创新者、使用者及广大公众之间形成互动协作格局,使技术监管从“控制与被控制”转向“共治共享”,集多方合力提升技术监管质效[19]。

(三)监管介入的影响因素与决策机制

监管因何介入、何时介入影响着监管行动与技术演进进程的耦合适配性,是技术治理研究与实践的重要命题,其决策受到多重因素影响。一是主观认知水平。准确把握技术成熟度、扩散水平与外部性等是科学决策的前提[20]。监管者对技术涌现的捕捉、发展现状的把握与监管需求的感知深刻影响着介入决策判断。二是客观条件约束。一方面,监管资源具有有限性,监管者需统筹监管安排以趋近帕累托最优[21];另一方面,既有制度若无法支撑新技术的监管需要,还需进行制度修订或新立工作,导致监管时点延后。三是多元主体博弈。尽管“决策黑箱”尚未完全消解,但现代决策系统并非全然封闭[22],科技企业、学术机构、社会组织及广大公众提出的不同利益诉求均对监管决策产生影响[23]。四是国际竞合影响。随着全球技术合作的深化及技术外部性的增强,国际层面的竞争张力与协调需求也对各国本土的监管安排带来影响。五是风险事件冲击。以颠覆性技术为代表的技术创新极易引发高影响性的风险事件,迫使监管者在压力情境下进行抉择[24]。

整体而言,监管决策机制可以被视为一个“监管缺口识别—监管议题输入—监管策略制定—监管决策输出”的动态系统,如图 1 所示。一般情况下,其逻辑起点在于新技术涌现之后的监管需求觉察,随即生成决策议题、启动决策程序。在多重因素影响下,监管者或是保持观望,或是择时介入,进而对技术产生抑制或促进作用,实现鼓励或抑制创新、防范风险等监管目标。

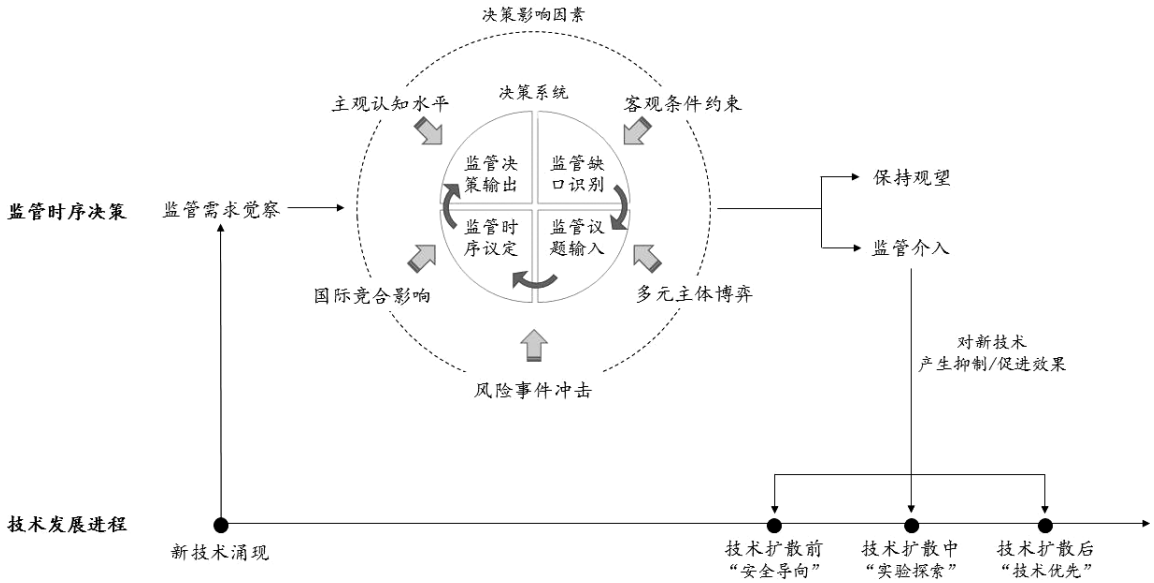

图 1　一般情况下技术监管介入的影响因素与决策机制

二、探寻超越“预防”与“放任”的第三条治理之道

面对 AI 治理中“过早扼杀创新”与“过晚风险失控”的两难困境,“预防优先”与“市场自由”等既有范式均显局限。科林格里奇悖论所揭示出的监管时机困境,本质上源于线性治理思维与非线性技术演进的错配。因此,急需突破二元框架,构建兼顾时序灵活性与强度适配性的动态治理新范式。

(一)传统技术治理范式的局限性与 AI 治理的新挑战

AI 技术已展现出显著区别于既往技术的发展态势,使 AI 治理迎来重大挑战,尤其是在“时”与“度”两个关键方面。从“时”的方面来看,传统技术治理范式以技术扩散前的预防节点和扩散后的调节节点为主流介入时机,但当前 AI 技术的非线性演化态势正以远高于既往技术的频率扰动着经济社会秩序与监管安排[25],令锚定某个特定节点介入的应对之策已显不适。例如,欧盟本意在 AI 技术扩散前推出《人工智能法案》,却在立法过程中就受到 ChatGPT 等生成式 AI 产品问世与迭代更新的多次冲击;美国将 AI 技术发展速度置于优先地位,现今却难以有效化解算法黑箱、算法偏见、数据安全等沉积问题。从“度”的方面来看,一方面,传统技术治理范式有相对明确的边界,而 AI 模型能够跨主体、跨任务、跨场景渗透至千行百业,这一技术层面的底层通用性特质使 AI 治理需要面向更多的治理对象和更广的治理范围统筹治理强度[26];另一方面,较既往技术而言,AI 技术的不可解释性、难以预测性、演变动态性和影响广泛性都更加显著,在治理强度基准设定方面已不能简单套用既有方法[27]。此外,AI 所依赖的数据、人才等核心资源可流动性强、迁移成本低,在监管套利驱动下,不同国家和地区间还可能出现逐底竞争,抬升 AI 治理强度的失准风险。

(二)“适时性”治理:基于社会外部性的动态触发机制

监管介入时序的选择,核心在于权衡“控制成本”与“信息完备性”之间的张力关系[28]。传统思路往往主张在技术研发早期就寻求干预支点,然而,面对 AI 技术非线性、涌现性演化特征[29],这种建立在“技术决定论”之上的线性治理逻辑愈发难以适用。因此,有效的治理范式应当实现监管重心的视阈转换,即从聚焦实验室内部的“技术生成逻辑”,转向聚焦技术进入社会系统之后的“外部性显现逻辑”[30]。换言之,政府监管的合法性边界与介入时机,不宜简单以技术成熟度(TRL)为唯一参照[31],而应更多依据该项技术在社会场域中的社会渗透率(SPR)及其所引致的风险暴露程度来加以界定[1]。基于这一视角,本文提出构建以 SPR 阈值为基础的“阶梯式滚动触发机制”,摒弃静态的时间预设,转而以技术应用的广度(用户规模)与深度(场景敏感性)为动态坐标,构建 3 个递进的治理区间,以实现监管供给与技术扩散、风险演进的动态匹配[32]。首先,在技术发展的潜伏期,治理应当遵循以“备案监测”为主的审慎观察逻辑。当技术产品刚刚完成初步商业化部署,其社会渗透率尚处低位(如仅在小众群体或封闭测试环境中扩散),且未切入关键基础设施领域时,尽管技术发展的不确定性最高,但其社会负外部性尚未显性化。此时若引入过早的强力规制,极易引发“寒蝉效应”,阻断技术创新路径[33]。因此,监管应恪守“最小干预原则”,将重心置于信息获取而非行为矫正。在政策工具的选择上,可实施“告知性备案”制度,要求技术供给方报备算法类型、基本逻辑及预设应用场景[34]。建立“监管雷达”以保持对技术迭代的敏捷感知,同时不设置实质性的准入门槛,从而为技术试错与商业模式的早期探索预留充足的弹性空间。其次,随着技术进入扩散期,治理重心应转向以“信息披露”为核心的规制响应。当社会渗透率跨越关键临界点,或技术开始向金融、教育、舆论生成等具有公共属性的垂直领域渗透,信息不对称所引致的算法偏见、误导性决策等风险开始具象化[35]。此时,监管目标需从单纯的“保护创新”转向“矫正失灵”,其核心在于打破算法黑箱[36],通过增强透明度来保障用户的知情权与选择权。相应的政策工具也应升级为包括要求服务商披露算法归因逻辑与训练

数据来源、推行 AI 生成内容标识机制[37],并确立用户的“退出权”与“解释权”的强制透明度监管[38]。通过强制披露缓解技术与公众间的信息势差,利用市场选择机制倒逼企业进行算法伦理优化。最后,当技术演化至规制期,治理则应当必须升级为以“许可审计”为重的底线管控。一旦 AI 技术演化为社会基础设施,具有极高的社会渗透率和强锁定效应,或广泛介入生育养老、医疗诊疗、自动驾驶等涉及生命安全与公共秩序的高风险场景,其风险便具备了系统性与不可逆性[39]。此时,单一的市场自律与透明度机制面临失效,政府作为公共利益的最终守护者必须履行公共安全兜底责任,将监管强度提升至最高层级以贯彻预防原则[40]。在此阶段,应实施“事前许可”与“全流程审计”制度,建立严格的算法安全评估准入标准,在关键决策环节引入“人在回路”的强制干预机制[41],并构建可回溯的责任分担体系。对于可能引发系统性危机的特定技术应用,监管机构甚至应保留实施“熔断”与“禁入”的最终规制权力,以确保社会公共利益的底线安全[42]。

(三)“适度性”治理:基于比例原则的监管强度动态校准机制

“适时性”治理解决了监管介入的时序问题,但并未完全回答介入的强度与方式应如何确定。面对 AI 技术应用场景的无限多样性及其风险谱系的极大异质性,对其治理亦非易事,治理强度的确立无法沿用均质化的单一标准,而需遵循行政法比例原则,在监管手段的侵入性与所规制风险的破坏性之间构建一种动态均衡[43]。这种均衡的实现,依赖于从“技术本位”向“场景本位”的逻辑转向,以及从“静态规制”向“动态校准”的机制演进[44-45]。一是风险的场景化耦合与治理工具的级差配置。AI 技术的风险本质,并非单纯内生于算法代码或数据本身的静态属性,而是技术特征与具体应用场景发生耦合时的涌现性产物。同一套 AI 大语言模型,嵌入娱乐对话场景与医疗诊疗场景,其社会外部性存在天壤之别。因此,治理强度的基准设定自然应当超越单一的技术本体维度,转而构建以“危害后果严重性”与“风险发生概率”为核心坐标的双维评估框架。基于此,治理工具的供给应当呈现出与风险等级严密咬合的阶梯式结构,而非扁平化的普适性规则。在这一光谱的顶端,针对触及伦理底线或可能引致系统性危机的“不可接受风险”场域(如大规模生物特征监控),治理机制表现为绝对的刚性,即通过法律层面的“禁入”阻断风险源头;随着风险等级下沉至涉及公共安全的“高风险”场域,治理逻辑转向“预防性合规”,通过事前许可与全流程审计强化责任约束;而在广泛存在的“有限风险”与“低风险”场域,治理强度则应显著回撤,转而依赖透明度机制赋权用户,或退守至行业自律与软法引导层面。此种“金字塔”式的级差配置,旨在确保行政权力介入始终被锁定在必要的限度之内,避免过度规制对技术微创新的无差别挤出[46]。二是成本收益的动态权衡与监管强度的双向调节。更进一步而言,适度性治理的本质在于其动态适应性,由此衍生而来的治理并非一套静态的制度安排,而是一个必须随技术本身迭代而持续演进与优化的过程。面对 AI 技术迭代带来的不确定性,治理强度的维持需要引入严格的成本—收益分析范式[47]。当某项监管措施的边际社会成本(包含合规负担与抑制创新的机会成本)超过其带来的边际风险防范收益时,该治理结构便处于失衡状态。因此,理想的治理范式内部应当内嵌一套灵敏的、具备“双向调节”弹性的“强度校准回路”治理机制。一方面,随着特定技术路线的成熟或安全验证的通过,监管应当具备“降级”的通道,及时豁免不必要的审批义务以释放技术红利;另一方面,当技术应用发生场景异化或衍生出新型风险时,治理体系亦能迅速启动“升级”响应,实施熔断或提级管控。通过这种基于反馈的动态校准,治理强度得以在保障公共安全底线与最大化创新收益之间,持续逼近帕累托最优的均衡点[45]。

三、全球主要经济体 AI 监管的实践比较与理论反思

随着 AI 应用外溢至经济社会核心领域,主要经济体在监管介入的时序选择(适时性)与强度配置(适度性)上呈现出显著分化。基于监管介入的

时序选择与强度配置,各主要经济体的 AI 治理模式可归纳为预防导向型、市场驱动型、国家主导型和混合平衡型等典型范式。系统梳理并深入比较欧、美、英、中的实践图谱,为理解“适时性”与“适度性”的制度逻辑提供了丰富的观察样本,有助于检验两个核心维度在现实中的表现形式与作用机制,并为构建科学合理的 AI 治理体系提供坚实的实证支撑、深刻的理论指导及切实可行的路径借鉴。

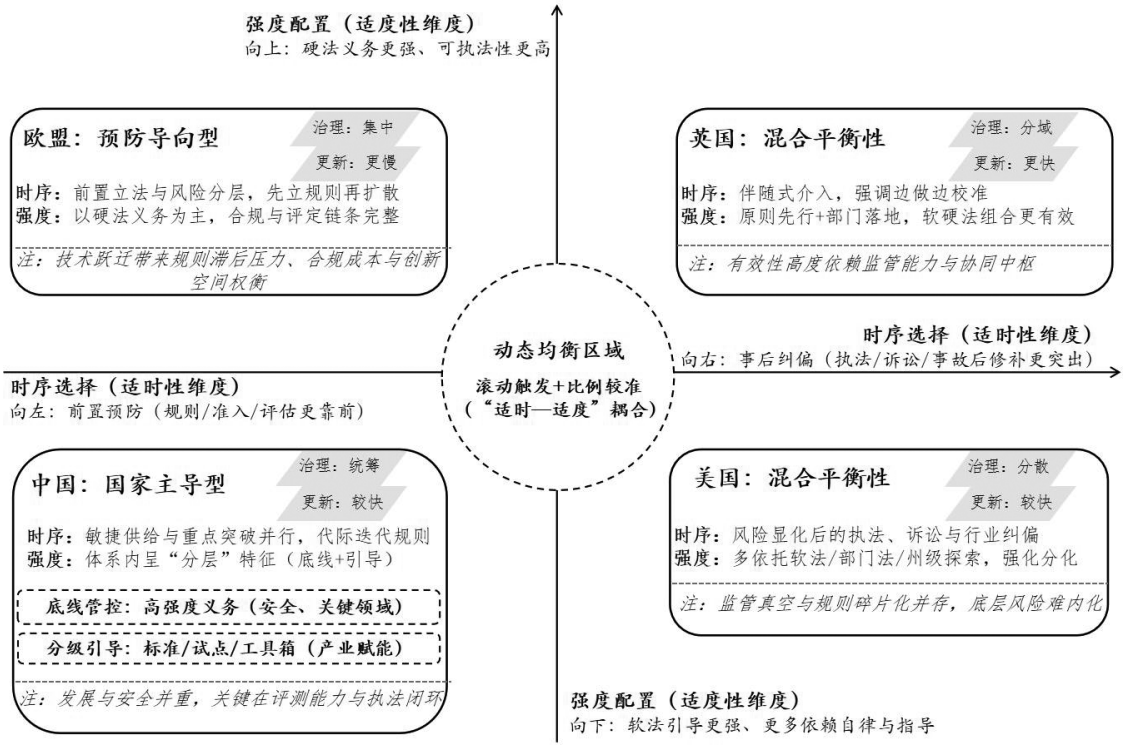

图 2　全球主要经济体 AI 监管范式比较框架

资料来源:作者根据欧盟 AI 法案、英国 AI 白皮书、美国相关政策框架及中国算法/生成式 AI 监管文件整理。

(一)欧盟:预防导向型下的“时序抢跑”与“强度固化”

欧盟 AI 监管的核心特征是试图通过立法技术的完备性消除技术演进的不确定性。在“时序选择”维度,采用“抢跑式介入”。从 2021 年 4 月提议至 2024 年通过的《人工智能法案》标志首部系统性 AI 立法诞生[42],早于技术大规模扩散。该策略根植三重逻辑:其一,延续《通用数据保护条例》(GDPR)权利优先于效率的监管传统及“预防原则”;其二,通过早期统一立法规避成员国规则碎片化;其三,利用“布鲁塞尔效应”以规则优势弥补产业劣势[48]。然而,这种“抢跑式介入”面临预防性立法的固有困厄。立法进程耗时 3 年,而生成式 AI(如 ChatGPT)的爆发性问世迫使法案仓促增补 GPAI 条款,要求高风险模型进行安全评估与透明报告,暴露了静态法律文本在应对技术跃迁时的滞后性。

在“强度配置”维度,欧盟《人工智能法案》呈现出较高“颗粒度”的风险分层逻辑。法案将 AI 系统分为禁止类(如社会评分)、高风险类(如关键基础设施、执法等)、透明度义务类及最小风险类 4 个层级,并相应配置了从绝对禁止到事前合格评定的差异化规制手段[49]。然而,“高颗粒度分层”在实践中面临两重挑战。一是场景漂移与分类粘滞。通用目的模型在不同下游场景(如娱乐 vs 医疗)中的风险属性差异显著,虽然法案试图以“预期用途”为锚点,但在技术快速迭代下,边界维护与执法识别成本依然高昂[42]。二是合规基准线整体偏高。高风险系统的一揽子事前合规义务虽能构筑安全屏障,却显著增加了企业成本,可能挤压中小企业的创新空间并导致市场结构集中化[50]。

(二)美国:市场驱动型下的“时序滞后”与“强度分化”

美国 AI 监管模式是典型的市场驱动下自由主义治理,其核心理念是通过延迟干预预留“0 到 1”创新空间,依托市场机制平衡风险与收益。在“时序选择”维度,奉行“市场先行、监管跟进”。联邦层面未出台统一立法,初期治理主要依赖行业自律及《人工智能权利法案蓝图》(2022)等非约束性文件。特朗普重新执政后,2025 年相继发布《消除美国 AI 领导力障碍行政命令》《美国 AI 行动计划》,废除既往安全限制、强化基建投资、松绑监管。同年 12 月出台的《确保国家人工智能政策框架》更确立了“最小负担”联邦标准,旨在遏制州级立法碎片化并收回监管主导权,深刻体现了其“滞后介入”的政治法律传统。其一,美国宪法确立的联邦制结构使技术监管权限分散于联邦与州两级,国会层面既缺乏统一 AI 立法的政治共识,也缺乏明确授权,行政部门多通过既有行业监管框架间接规制 AI 应用[51-52];其二,“技术中立”与“市场自律”的自由主义传统深刻塑造监管理念,政策制定者普遍信赖市场竞争与司法事后救济纠偏技术风险,对事前强干预保持高度警惕;其三,在全球 AI 竞争中,美国将“宽松、克制”的监管视为吸引人才与企业、巩固技术优势的战略工具。然而,这种滞后介入策略正面临州际规则冲突、监管空白与系统性风险累积的持续考验[53]。在联邦立法缺位下,各州通过差异化立法填补空白,如加州《人工智能系统透明度法案》、纽约州《负责任人工智能安全与教育法案》等,但由此导致企业合规义务高度碎片化,跨州运营成本显著上升。长远看,“监管滞后”虽为技术试错预留了空间,却使算法黑箱、算法偏见与数据安全等问题持续累积[54]。当风险显性化时(如算法导致的大规模就业歧视诉讼、自动驾驶系统致命事故),监管者往往陷入“亡羊补牢”的被动境地,纠偏成本与社会代价高昂。

在“强度配置”维度,美国采取“场景驱动、工具灵活”的行业本位模式。由于缺乏统一的风险分层立法,监管强度主要嵌入既有部门监管框架。在金融服务领域,联邦贸易委员会(FTC)与消费者金融保护局(CFPB)依据《公平信用报告法》《平等信用机会法》等,对 AI 信用评分和贷款审批算法实施较严的可解释性与反歧视要求;在医疗领域,药品监督管理局(FDA)将 AI 医疗器械纳入既有医疗器械分级审批体系;在自动驾驶领域,国家公路交通安全管理局(NHTSA)多以指导文件而非法规约束,将安全验证责任主要交由制造商自我合规。这种“场景驱动”监管模式在初创企业活跃度、风险投资规模和技术迭代速度等方面显著激励创新,印证了“宽松环境促进创新”的政策预期,但也暴露出系统性监管能力不足。其一,横向协调碎片化。不同行业监管机构各自为政,缺乏统一的 AI 风险评估方法、透明度标准和责任分配规则,跨领域应用面临多头监管与标准冲突。其二,纵向穿透不足。传统部门监管难以穿透大模型等通用技术,形成“九龙治水而无总责”监管盲区。其三,执法强度整体偏弱。在缺乏明确法律授权下,监管主要依赖指导文件、最佳实践等软法及事后调查与诉讼,事前预防与持续监督机制薄弱,在算法透明度、数据隐私等方面留下明显制度缺口,已引发大量消费者权益侵害与公平性争议[55]。

(三)英国:实用主义平衡下的“时序伴随”与“强度分域”

英国 AI 监管特征为“原则先行、能力赋能、分域执行”,通过灵活制度与敏捷监管,同步推进监管者能力建设与技术演进,实现“监管与创新共生”,在避免真空与过度干预间寻求动态均衡[56]。在“时序选择”维度,采取“伴随式介入”策略。现阶段政府未制定专门立法,而是通过 2023 年《一种支持创新的人工智能监管方法》白皮书确立五项跨领域原则(安全稳健、透明可解释、公平、问责与治理、可申诉与救济),由既有监管机构在各自职权内落地实施[57]。2025 年《AI 机遇行动计划》进一步将安全评测、算力资源与国际协作纳入统一框架,标志着治理体系从概念向实施的转型。这种“伴随式介入”时序选择有其深刻的制度与文化根源。其一,英国作为传统的普通法系国家,监管理念强调“案例积累”与“渐进演进”而非“成文

法系统构建”;其二,脱欧后英国试图以“监管灵活性”形成相对于欧盟的比较优势,吸引全球 AI 企业与投资[58];其三,FCA 等既有机构已积累丰富专业经验,“赋能现有机构”可最大化利用存量资源、降低制度转型成本[59]。然而,该模式面临监管能力与技术演进非对称竞争的挑战。其有效性高度依赖监管机构知识更新速度、跨部门协同效能及资源保障水平;当 AI 出现非线性突破时,该模式可能从“同步演进”退化为“被动追赶”。

在强度配置上,英国展现“原则统领、分域细化”的差异化结构,授权各行业监管机构在统一原则框架下自主设定执法强度,形成“适应性分层”体系。该模式的运作基于两个关键机制。一是监管责任纵向下沉。由专业机构根据场景风险设定强度“刻度”,精准匹配资源与风险。二是原则渐进转化。各机构通过行业指引、监管沙盒等将宏观原则转化为可操作标准,使监管强度随技术演进动态调整。然而,这种“原则统领—分域细化”的强度安排在提升制度弹性的同时也存在三类结构性短板。第一,法律约束力与合规可预期性偏弱。由于未采取统一立法,跨领域原则主要由机构解释适用,难以直接生成一体适用的硬性义务,易导致“原则一致、标准各异”。第二,跨域外部性与基础模型的通用性风险难以由单一行业监管者内化,易在监管边界处形成“空隙”或“多头治理”。第三,治理绩效高度依赖监管能力与资源供给。原则导向只有在评测工具、数据接口、专业队伍与跨部门协同机制持续跟进时,才能转化为可执行、可问责的强度刻度,否则极易滑向“软约束”。

(四)中国:国家主导型下的“时序双轨”与“强度分层”

中国在 AI 治理领域探索出了一条具有鲜明主体性特征的道路,其核心逻辑在于通过“国家统筹”实现发展与安全的辩证统一,构建了一种“敏捷立法、双轨驱动”的治理范式。在“时序选择”维度,中国实践了“急用先行”与“长远规划”相结合的双轨策略。中国采取“重点突破、分域推进”做法,以“小步快跑、迭代出台”的敏捷立法路径回应技术应用的迫切需求[56]。从 2017 年《新一代人工智能发展规划》奠定战略基础,到 2021 年的《算法推荐规定》和 2022 年的《深伪规定》,再到 2023 年的《生成式 AI 办法》,监管规则紧随技术代际更迭而快速迭代。这种“回应式立法”极大缩短了制度供给的时滞,为在技术风险暴露的早期迅速确立规则边界提供了可能。同时,通过“暂行办法”等形式预留制度接口,为后续根据技术演进调整规则保留了充裕的弹性空间。2025 年发布的《国务院关于深入实施“人工智能 +”行动的意见》进一步对政策供给进行了优化,较好地兼顾连续性与敏捷性,有效平衡了法律稳定性与技术变动性之间的张力。

在“强度配置”维度,中国构建了“底线管控 + 分级引导”的分层治理体系。一方面,在涉及意识形态安全、社会动员能力及关键信息基础设施的高风险领域(如生成式内容服务、舆论引导算法),实施严格的“双新评估”(安全评估与算法备案)制度,确立了不可逾越的红线;另一方面,在产业赋能、科学研究及企业级服务等垂直应用领域,通过《国家人工智能产业综合标准化体系建设指南》等软法工具及行业标准进行引导,大幅降低合规成本,预留和释放出充足的试错空间。这种“抓大放小、松紧有度”的配置逻辑,在确保国家安全底线不被击穿的前提下,最大程度激发了国产大模型的产业竞争力与技术创新活力[60],巧妙地将政治安全、社会稳定等刚性底线与产业发展的弹性空间结合起来,为全球 AI 治理贡献了“发展与安全并重”的中国方案。

(五)实践经验的理论反思与启示

通过深入比较分析欧、美、英、中典型 AI 监管模式,可清晰观察到“适时适度”抉择在不同治理语境下的多元实践形态及其深层规律。各经济体监管模式的显著分化根植于制度传统、发展阶段、价值取向的客观差异,深刻反映了面对 AI 技术冲击时传统治理范式的适应性重构与创新探索。一是“适时性”重在匹配技术扩散节律,而非简单的早或晚。欧盟以立法前置降低未来外溢风险,却承受制度时滞与规则粘滞;美国以市场先行释放试错空间,但在联邦缺位与州际差异中累积协调

成本;英国强调伴随式校准,对监管学习速度与协同效能提出更高要求;中国以双轨供给在早期确立底线并预留迭代接口。可见,“适时”的关键在于形成可持续更新机制,使监管随外部性显现与扩散程度动态调整[61]。二是“适度性”不等同于强监管或弱监管,而是强度刻度、合规成本与可执法性之间的比例性校准[62]。欧盟以高颗粒度风险分层与合格评定制度构筑强前置屏障,提升了权利保障与安全可验证性[63],但较高的合规基准线也抬升了进入门槛并可能强化市场集中。美国以场景化监管实现强度分化,创新激励更为显著,但横向协调不足与底层通用模型风险难以内化,使强度配置易出现结构性缺口。英国通过共同原则确立底线,再由行业监管者细化强度刻度,提升了场景适配性,却也出现标准不一与合规预期不稳定的摩擦。中国以底线管控与分级引导并行,在高风险领域形成高强度约束,在产业赋能领域保留较大政策弹性,但其有效性同样依赖评测能力、执法闭环与规则迭代的持续供给。可见,“适度”的实质是将强度选择嵌入可执行的责任链条与评估反馈机制之中,使监管既能承担风险兜底责任,又不过度挤压创新与竞争结构[64]。三是治理模式和结构决定监管绩效的上限,硬法与软法的组合方式须与国家能力相匹配[65]。欧盟以成文法与统一市场逻辑塑造规则稳定性,适于建立跨国一致性预期,但需要以更强的二级规则与执行机制来对冲技术跃迁带来的制度滞后。美国的联邦制与自由主义传统强化了分布式试验与市场纠偏,却必须通过最低联邦基线或更强的协调机制来降低碎片化成本。英国的普通法渐进路径与能力型监管凸显灵活优势,但其制度有效性高度依赖监管资源投入与跨部门协调中枢。中国的国家主导与统筹机制强化了快速动员与规则供给能力,但需要在标准互认、透明度工具与国际对接层面持续提升规则可迁移性。4 种经验共同表明,制度设计不仅是理念选择,更是能力工程,缺乏评测工具、数据接口与执法资源的制度安排,很难把原则性目标转化为可问责的监管实践。四是全球 AI 治理层面需要在多样性中寻求最低共识,避免监管竞争导致安全底线被稀释[66]。AI 风险具有跨境传播与链式扩散特征,单一经济体的规则难以独立闭合风险外部性。未来更可行的合作路径,不在于追求一体化立法,而在于围绕关键环节形成可操作的兼容性安排[67]。只有在承认各国制度差异与发展阶段差别的基础上构建可协商、可互认、可迭代的国际协调网络,才能在竞争格局中维持必要的安全阈值,并为 AI 向善与普惠发展提供更稳定的制度环境[68]。

四、迈向“适时—适度”耦合治理,构筑敏捷审慎人工智能监管体系的路径

当前,AI 指数级迭代并深度嵌入经济社会各领域,传统规则前置或事后追责监管难以应对。急需转向动态治理能力建设,依据技术成熟度与风险外部性构建可触发介入与可校准强度机制,重点推进时序触发、强度校准、主体协同、空间协调、机制迭代五方面,形成监测预警、分级合规、协同共治、国际互认的评估反馈衔接路径,如图 3 所示。

(一)锚定社会渗透率阈值,构建“三期三策”滚动触发机制

实现监管介入与技术扩散及外部性显现的同步,是“适时性”治理的关键。可将 SPR 与场景敏感性作为操作性锚点,构建潜伏期、扩散期、规制期相衔接的“三期三策”滚动触发机制。一是量化 SPR 阈值并建立预警体系。建立国家级 AI 发展与风险监测平台,构建综合用户规模、使用频次与场景敏感度的 SPR 测算模型。在潜伏期(SPR 较低且未进入关键基础设施),以告知性备案和“监管雷达”为主,仅要求披露基本信息,为技术试错保留空间。在扩散期(SPR 跨越阈值或进入金融、教育、舆论等公共领域),自动触发强制透明度监管与 AI 生成内容标识制度,赋予用户知情、解释与退出权。在规制期(SPR 高度集中或嵌入医疗、自动驾驶等高风险场景),实施事前许可与全流程审计,引入“人在回路”机制并保留熔断和禁入权。二是建立监管强度升降级通道。潜伏期技术设定有限观察期,期满无实质风险可减轻或取消管制,防止监管惯性固化;扩散期和规制期技术引入年度评估,根据风险事件、合规表现和安全能力建设

动态调整监管级别,避免“只升不降”的棘轮效应。三是完善情景驱动的前瞻预案。面向通用大模型、具身智能、脑机接口等前沿技术,预设演化情景并配套政策工具组合,在能力跃迁或意外风险出现时快速切换预案,减少监管反应的滞后性和被动性。

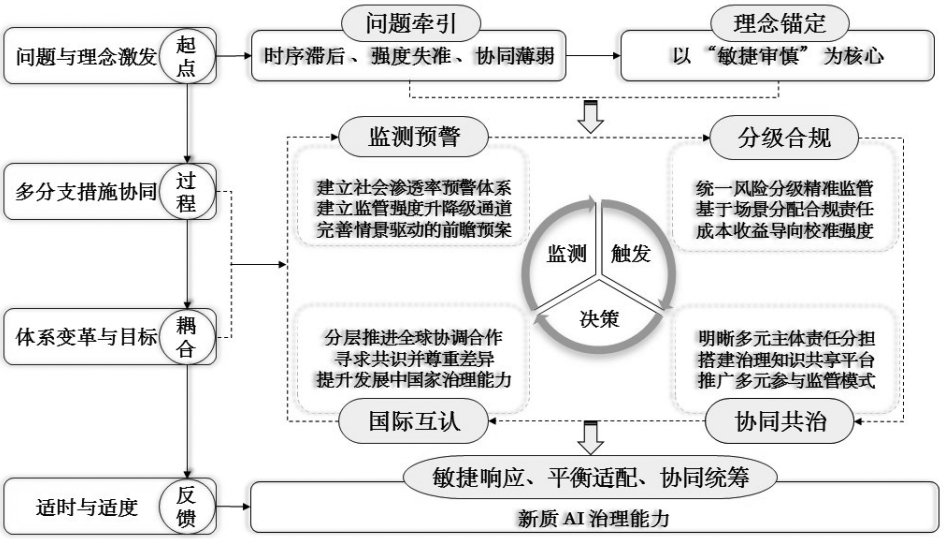

图3　构筑前瞻柔性人工智能监管体系的可行路径

(二)匹配风险比例原则,建立“四级四策”精准监管工具箱

“适度性”治理要求以比例原则动态校准监管强度,并配置从“禁入”到“自律”的阶梯式工具。第一,构建统一的风险分级与工具对应关系。将AI应用划分为不可接受风险、高风险、有限风险和低风险四个等级。不可接受风险(如大规模社会评分、隐性操纵)实行绝对禁止;高风险领域(医疗诊断、司法裁判、关键基础设施)实施事前许可、周期审计和强制责任保险;有限风险场景(教育推荐、就业筛选)强化透明度披露、用户知情同意和算法影响评估;低风险应用(娱乐内容生成、一般个性化推荐)依托行业自律和软法引导,仅需标注AI生成属性。第二,细化通用模型的场景化责任分配。对大规模预训练基础模型实施“基础安全评估+场景责任分担”的双重机制。模型开发者完成能力边界说明、系统性风险评估和红队测试等基础义务,纳入基础模型安全数据库。下游开发者根据具体用途承担与场景等级相称的合规责任,监管机构据此动态调整强度,避免“一刀切”或监管真空。第三,建立成本收益导向的强度校准机制。设立专门的AI治理效能评估机构或委员会,定期对高风险类别的监管绩效进行评估,比较合规成本、创新影响和风险防范成效。当边际成本明显超过边际收益时,依法启动简化程序或部分义务豁免。当出现新型风险或重大事故时,及时提高强度。监管沙盒可作为强度调整的试验场,为降级或升级提供实证支撑。

(三)激活协同治理网络,构建多元协同治理格局

面对技术的复杂性,监管亟待从政府“单中心管制”转型为政府、企业、学界与社会组织的“开放协同网络”。其一要明晰多元主体的责任分工与协同机制。政府侧重制度供给、标准制定、执法监督与能力建设,企业作为责任主体承担伦理审查、风险自评估与信息披露义务,学术机构与专业第三方提供独立评测、技术审查和政策咨询,公众与社会组织通过参与听证、舆论监督和公益诉讼等方式强化外部约束。可在国家或区域层面设立AI治理协调平台,形成常态化的议题协商、规则共创和反馈修正机制。其二要建设面向多方的治理知识共享平台。整合监管部门案例库、企业风险报告、研究机构评估结果和用户投诉反馈,构建覆盖主要领域的“风险地图”和“最佳实践库”。在做好

数据脱敏与商业秘密保护的前提下,分级开放相关信息,为企业合规决策和政策优化提供依据,同时提升社会整体的风险感知水平。其三要推广监管沙盒与“陪伴式监管”模式。将沙盒机制由金融扩展至医疗、教育、自动驾驶等重点领域,允许在可控范围内进行有限试点。监管机构适度嵌入研发和测试过程,提供实时合规指导和风险提示,将成功经验转化为可复制的规则模板,将失败案例形成负面清单向行业公开,借此在可控环境中实现制度与技术的双向学习。

(四)推动国际规则互认,在多样性中寻求最低共识

AI 风险具有显著跨境溢出效应,单一国家的监管难以完全内化外部性。现实中,监管竞争与规则分化并存,既可能推动制度创新,也可能诱发“逐底竞争”。因此,有必要在尊重差异的基础上,推动形成柔性的国际共识框架。首先,关键在于分层推进全球协调与合作。可在全球层面推动围绕若干底线问题形成原则性共识,如禁止 AI 服务于大规模杀伤性行为、严控滥用生物特征监控和严重歧视性算法。在区域与双多边层面,重点推动风险评估方法、透明度披露要求和测试标准的互认,降低跨境运营的合规摩擦。在具体风险事件层面,探索建立跨境协同应对机制,提升对重大数据泄露、系统性金融风险等事件的联合处置能力。其次,将合作重点放在程序互认而非实体完全同一。应优先推动评估报告格式、测试流程和披露框架的可比性与可转换性,以“最低标准加差异容忍”的方式,为各国保留必要的政策空间。最后,长远基础在于建设面向发展中国家的治理能力支持网络。通过提供评测工具、标准文本、培训课程和技术援助等方式,缩小不同国家监管能力的差距,避免监管洼地成为高风险活动的避难所。支持南南合作平台在 AI 治理经验和技术标准上的横向交流,形成更为均衡的全球治理能力结构。

(五)嵌入动态反馈回路,确保治理体系持续优化进化

治理体系需发展出与技术迭代同步的代谢能力,通过制度化的反馈循环,确保政策不过时、不僵化,实现规则的动态适配与持续进化。在工具与数据支撑层面,应建设面向治理绩效的综合监测与评估系统。通过统一的“监管仪表盘”,持续跟踪技术发展速度、风险事件态势、合规成本负担和国际协调进展等关键指标,一旦某一维度明显偏离预期,及时启动政策评估与规则调整程序。在规则供给与更新机制层面,有必要完善规范的有期限适用和快速修订安排。在重要监管文件中适度引入有效期审查和定期评估条款,到期后依据实际实施效果决定延续、修订或废止,防止过时规则长期占据制度空间。同时,在新技术领域建立简化程序和暂行性规范通道,缩短从问题识别到规则供给之间的时间差。在制度创新与区域实践层面,还应充分发挥地方试点的“前哨”功能。鼓励有条件的地区围绕特定应用场景开展差异化监管试验,在确保底线安全的前提下探索更具弹性的监管方式。对实践证明有效且可复制的做法,通过制度化程序上升为国家层面的规则工具;对具有积极探索意义的创新尝试,则给予适度容错与政策激励,逐步形成“地方先行、国家整合、国际对接”的治理升级路径。

参考文献:

[1] BALDWIN R, CAVE M, LODGE M. Understanding regulation: theory, strategy, and practice [M]. 2nd ed. Oxford: Oxford University Press, 2012.

[2]TAEIHAGH A. Governance of artificial intelligence [J]. Policy and society, 2021, 40(2): 137-157.

[3]曾婧婧,钟书华. 论科技治理[J]. 科学·经济·社会, 2011,29(1):113-118.

[4]苏成,赵志耘,赵筱媛,等. 颠覆性技术新阐释:概念、内涵及特征[J]. 情报学报,2021,40(12):1253-1262.

[5]樊春良. 国家战略科技力量的演进:世界与中国[J]. 中国科学院院刊,2021,36(5):533-543.

[6]贺德方,陈宝明,汤富强. 科技治理体系演变趋势与对策研究[J]. 科学学研究,2023,41(6):989-997.

[7]王海丹,伍浩松,王政. 国外主要有核国家核安全监管和法规体系概况及启示[J]. 中国核工业,2016(10):28-31.

[8]闫绪娴,侯光明,闫绪奇. 美国政府在科技发展中的作用及其对我国的启示[J]. 中国科技论坛,2004(3):130-133.

[9]董洁,孟潇,张素娟,等. 以色列科技创新体系对中国创新发展的启示[J]. 科技管理研究,2020,40(24):1-12.
[10]陈强. 德国科技创新体系的治理特征及实践启示[J]. 社会科学,2015(8):14-20.
[11]邱举良,方晓东. 建设独立自主的国家科技创新体系:法国成为世界科技强国的路径[J]. 中国科学院院刊,2018,33(5):493-501.
[12]程如烟,孙浩林. 主要经济体支持颠覆性技术创新的政策措施研究[J]. 情报学报,2021,40(12):1263-1270.
[13]周仲飞,李敬伟. 金融科技背景下金融监管范式的转变[J]. 法学研究,2018,40(5):3-19.
[14]范月蕾,王慧媛,姚远,等. 趋势观察:生命科学领域伦理治理现状与趋势[J]. 中国科学院院刊,2021,36(11):1381-1387.
[15]LIEBERT W, SCHMIDT J C. Collingridge's dilemma and technoscience [J]. Poiesis & Praxis, 2010, 7(1/2): 55-71.
[16]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.
[17]王健. 现代技术伦理规约的困境及其消解[J]. 华中科技大学学报(社会科学版),2006(4):82-87.
[18]张光君,彭池. 颠覆性技术的科林格里奇困境与合规治理出路[J]. 科技管理研究,2024,44(20):1-8.
[19]童云峰. 走出科林格里奇困境:生成式人工智能技术的动态规制[J]. 上海交通大学学报(哲学社会科学版),2024,32(8):53-67.
[20]辛竹琳,魏凤,邓阿妹,等. 基于技术成熟度的技术评价方法研究[J]. 科技管理研究,2024,44(11):80-89.
[21]孙志建. 怎样合理配置有限的政府监管资源:基于风险的监管模式的兴起及其潜在运行风险[J]. 上海行政学院学报,2022,23(2):32-44.
[22]刘红岩. 国内外社会参与程度与参与形式研究述评[J]. 中国行政管理,2012(7):121-125.
[23]王立,王峥,王永梅. 公共政策过程中的利益考量:基于利益相关者理论的分析[J]. 管理学刊,2012,25(4):80-84.
[24]刘一弘,钟开斌. 学习与竞争:重大突发事件如何触发政策变迁的文献述评[J]. 公共行政评论,2021,14(6):24-43,197.
[25]周佑勇. 论智能时代的技术逻辑与法律变革[J]. 东南大学学报(哲学社会科学版),2019,21(5):67-75,147,2.
[26]魏巍,曾铮,刘蕾. 从 DeepSeek 突破看我国人工智能产业创新范式、挑战与应对[J]. 经济纵横,2025(6):102-114.
[27]唐要家,唐春晖. 创新不确定性与人工智能监管创新[J]. 东北财经大学学报,2025(5):3-14.
[28]COLLINGRIDGE D. The social control of technology [M]. New York: St. Martin's Press, 1980: 19-21.
[29]RAHWAN I, CEBRIAN M, OBRADOVICH N, et al. Machine behaviour [J]. Nature, 2019, 568(7753): 477-486.
[30]姜李丹,薛澜. 我国新一代人工智能治理的时代挑战与范式变革[J]. 公共管理学报,2022,19(2):1-11,164.
[31]MANKINS J C. Technology readiness levels: a white paper[R]. Washington, DC: NASA, 1995.
[32]World Economic Forum. Agile governance: reimagining policy-making in the fourth industrial revolution[R]. Geneva: WEF, 2018.
[33]SCHAUER F. Fear, risk and the first amendment: unraveling the chilling effect [J]. Boston University Law Review, 1978, 58(5): 685-732.
[34]European Commission. White paper on artificial intelligence: a European approach to excellence and trust [R]. Brussels: European Commission, 2020.
[35]O'NEIL C. Weapons of math destruction: how big data increases inequality and threatens democracy [M]. New York: Crown, 2016.
[36]PASQUALE F. The black box society: the secret algorithms that control money and information [M]. Cambridge: Harvard University Press, 2015.
[37]国家互联网信息办公室,中华人民共和国工业和信息化部,中华人民共和国公安部. 互联网信息服务深度合成管理规定[EB/OL]. (2022-12-11)[2025-12-18]. https://www.cac.gov.cn/2022-12/11/c_1672221949354811.htm.
[38]WACHTER S, MITTELSTADT B, FLORIDI L. Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation [J]. International data privacy law, 2017, 7(2): 76-99.
[39]BOSTROM N. Superintelligence: paths, dangers, strategies [M]. Oxford: Oxford University Press, 2014.
[40]SUNSTEIN C. R. Laws of fear: beyond the precautionary principle [M]. Cambridge: Cambridge University Press, 2005.
[41]SHNEIDERMAN B. Human-centered AI [M]. Oxford: Oxford University Press, 2022.
[42]European Commission. Regulation (EU) 2024/1689 of the European Parliament and of the Councilof 13 June 2024 [EB/OL]. (2024-07-12) [2025-12-28]. https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng.
[43]蒋红珍. 比例原则适用的范式转型[J]. 中国社会科

学,2021(4):106-127,206-207.
[44]National Institute of Standards and Technology (NIST). AI risk management framework (AI RMF) playbook [EB/OL]. (2025-02-06) [2025-12-28]. https://www.nist.gov/itl/ai-risk-management-framework/nist-ai-rmf-playbook.
[45] OECD. Recommendation of the Council on Regulatory Policy and Governance[EB/OL]. (2012-11-02) [2025-12-28]. https://www.oecd.org/en/publications/recommendation-of-the-council-on-regulatory-policy-and-governance_9789264209022-en.html.
[46] AYRES I, BRAITHWAITE J. Responsive regulation: transcending the deregulation debate [M]. NewYork: Oxford University Press, 1992.
[47]Office of Management and Budget (OMB). Circular A-4: regulatory analysis[R]. Washington, DC: Executive Office of the President, 2023.
[48]皮勇. 欧盟《人工智能法》中的风险防控机制及对我国的镜鉴[J]. 比较法研究,2024(4):67-85.
[49]曾雄,梁正,张辉. 欧盟人工智能的规制路径及其对我国的启示:以《人工智能法案》为分析对象[J]. 电子政务,2022(9):63-72.
[50]European Commission. Impact assessment accompanying the Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts. [EB/OL]. (2021-04-21) [2025-12-28]. https://eur-lex.europa.eu/legal-content/EN/TXT/? uri = celex%3A52021SC0084.
[51]蔡翠红,张璐瑶. 全球人工智能治理探究:基于委托—代理理论视角[J]. 国际政治研究,2025,46(2):9-35,5.
[52]YOO S C, LAI A. Regulation of algorithmic tools in the United States[J]. Journal of law and economic regulation, 2020, 13(2): 7-22.
[53]李益斌,李浩洋. 欧美中人工智能监管规范比较研究[J]. 当代世界与社会主义,2024(5):161-169.
[54]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.
[55]刘兴华. 数字全球化时代的技术中立:幻象与现实[J]. 探索与争鸣,2022(12):34-44,210.
[56]司伟攀,徐峰,刘鑫怡. 全球人工智能立法实践与我国路径研究[J]. 情报杂志,2025,44(10):195-207,180.
[57]UK Department for Science, Innovation & Technology. A pro-innovation approach to AI regulation: government response [EB/OL]. (2024-02-06) [2025-12-25]. https://www.gov.uk/government/consultations/ai-regulation-a-pro-innovation-approach-policy-proposals/outcome/a-pro-innovation-approach-to-ai-regulation-government-response.
[58]周辉,金僖艾. 英国人工智能监管实践、创新与借鉴[J]. 数字法治,2023(5):194-206.
[59]吴钟灿,郝文强. 支持创新的人工智能监管:基于英国的经验借鉴与政策启示[J]. 图书与情报,2025(5):61-72.
[60]魏巍,刘蕾. 我国人工智能应用市场准入制度的权责划分、治理逻辑与路径重构[J]. 当代经济管理,2025,47(9):28-36.
[61]薛澜,赵静. 走向敏捷治理:新兴产业发展与监管模式探究[J]. 中国行政管理,2019(8):28-34.
[62]付鉴宇. 类 ChatGPT 大模型赋能数字政府的合比例性控制[J]. 科技与法律(中英文),2025(2):63-74.
[63]崔星璐,姚长青. 我国人工智能系统分级透明制度的构建路径:基于欧盟《人工智能法案》风险评估模式的研究[J]. 情报杂志,2025,44(11):160-169.
[64]张慧,李秋甫. 新兴科技的预防式伦理治理路径探析[J]. 自然辩证法研究,2024,40(2):96-103.
[65]张凌寒. 中国需要一部怎样的《人工智能法》?:中国人工智能立法的基本逻辑与制度架构[J]. 法律科学(西北政法大学学报),2024,42(3):3-17.
[66]贾开,赵静,傅宏宇. 应对不确定性挑战:算法敏捷治理的理论界定[J]. 图书情报知识,2023,40(1):35-44.
[67]李晓楠. 跨境人工智能服务安全监管的逻辑基础与规则展开[J]. 河北学刊,2025,45(5):170-180.
[68]俎文天. 人工智能全球治理合作:问题、进路与中国参与[J]. 国际经济评论,2025(4):153-176,8.