面向人工智能伦理风险的敏捷治理范式：国际经验与中国路径

顾建中[1]*，梁慧[1]

（1. 洛阳理工学院，河南洛阳）

摘要：人工智能技术的快速演进正持续放大其在公平性、透明度、责任归属与社会影响等方面的伦理风险。面对技术迭代周期远短于传统治理机制响应周期的现实矛盾，全球范围内正兴起一种以动态适配、多主体协同和迭代反馈为特征的敏捷治理范式。本文系统梳理该范式的历史演进逻辑，辨析其在原则嵌入、制度设计与实践落地三个维度的核心主张；对比分析不同区域在治理节奏、权责配置与技术适配路径上的结构性差异；深入探讨标准碎片化、能力鸿沟、价值张力与问责模糊等深层挑战；在此基础上，提出兼顾响应速度与制度韧性的中国路径构想------强调法治框架下的弹性授权机制、分层分类的风险响应体系，以及面向公共价值的技术能力建设。全文立足治理效能与伦理目标的双重实现，旨在为构建可持续的人工智能治理体系提供理论参照与实践思路。

关键词：人工智能伦理；敏捷治理；风险治理；技术治理；制度适应性

1. 引言：技术加速时代对治理范式的根本性挑战

1.1 技术迭代与制度滞后的结构性张力

人工智能技术呈现显著的指数级演进特征，其模型参数规模、训练数据量与推理能力迭代周期持续压缩至数月甚至数周量级；而传统监管规则制定平均耗时逾十八个月，修订程序受多重法定要件约束，刚性结构难以响应技术突变。这种技术加速与制度惯性的结构性张力，致使规制空白常态化、风险识别滞后化、治理适配碎片化，构成敏捷治理范式建构的根本动因 [1, 2]。

1.2 伦理风险的复合性与情境依赖性

人工智能伦理风险本质上具有高度的动态性与情境嵌入性，其表现形态随技术部署场景、社会文化语境及用户群体结构的变化而持续重构。医疗诊断系统中的偏见识别机制在金融信贷场景中可能失效，教育推荐算法在城乡差异显著的区域亦呈现非线性风险跃迁。这种复合性导致任何静态、去情境化的伦理准则均难以覆盖全谱系风险，治理设计必须内嵌实时反馈与场景适配能力 [2, 3]。

1.3 敏捷治理作为方法论转向的必然选择

敏捷治理绝非对规范约束的消解，而是以制度性反馈回路的结构性压缩为前提，通过提升政策学习速率、增强规则迭代频次与细化响应单元粒度，实现伦理原则与技术演进节奏之间的动态适配。其本质是将静态合规转化为持续校准，在不确定性中锚定价值稳定性 [4]。

2. 历史脉络与发展：从审慎监管到迭代共治的演进逻辑

2.1 早期技术治理的线性模型及其局限

传统技术治理长期依赖事前许可、标准先行与专家主导的线性模型，其制度设计预设技术演进路径可预测、风险边界可界定、影响范围可量化。然而，生成式人工智能展现出高度涌现性、跨域渗透性与实时演化性，导

致既有监管工具在响应时效、知识适配与责任归因维度出现系统性衰减。当模型行为无法被静态规则充分覆盖，当训练数据分布持续漂移，当部署场景呈指数级泛化，线性治理的解释力与执行力同步塌缩。这种结构性失配，本质上源于治理逻辑与技术本体论之间的根本张力 [2, 5]。

2.2 中间阶段的适应性尝试：沙盒机制与软法实践

监管沙盒与软法实践构成人工智能治理演进中的关键中间形态，其本质在于通过可控实验环境验证制度设计的适应性边界。行业指南与伦理宪章虽不具强制效力，却为技术主体提供可操作的行为锚点，在算法透明度、数据最小化等维度形成共识性规范。沙盒机制尤其凸显制度弹性测试功能，允许在真实场景中观测模型偏差演化路径与干预措施响应时滞，积累关于治理阈值设定的实证依据。此类过渡形态并非权宜之计，而是制度学习闭环的核心枢纽，持续校准硬法生成的时机与颗粒度 [5, 6]。

2.3 敏捷治理范式的成熟标志与共识凝聚

敏捷治理范式的成熟标志体现为制度性能力的结构性跃迁。多阶段评估机制突破传统线性审批逻辑，嵌入动态反馈回路；跨部门常设协调单元实现政策制定、技术部署与社会响应的实时对齐；实时数据驱动的风险监测网络依托标准化接口与语义互操作协议，支撑风险信号的毫秒级识别与分类分级。三者共同构成可测量、可验证、可迭代的治理基础设施，标志着治理主体从被动响应转向主动塑形，治理节奏从周期性调整转向连续性校准，治理依据从经验判断转向证据锚定 [7]。

3. 核心论点 A：原则嵌入的动态化------从静态声明到过程内化

3.1 伦理原则的可操作性转化瓶颈

伦理原则的可操作性转化面临双重结构性瓶颈：技术层面，公平性难以在模型训练中被形式化为可优化目标，因不同公平定义间存在数学互斥性；可解释性常受限于深度神经网络的内在黑箱特性，局部解释方法无法保障全局一致性；可控性则缺乏统一的干预接口标准，导致人工介入点模糊。组织层面，跨职能团队对原则的理解存在语义漂移，算法工程师、产品负责人与合规人员对"可控"的操作定义显著分化；开发流程未嵌入实时伦理评估节点，测试阶段才引入原则校验，形成滞后反馈闭环。更关键的是，现有工具链缺乏将抽象约束映射至具体超参数、损失函数项或部署阈值的标准化转换协议，致使原则悬浮于文档层而无法渗入代码层与决策流 [4]。

3.2 生命周期嵌入机制的设计逻辑

伦理考量的生命周期嵌入并非线性叠加，而是依据技术演进阶段的异质性特征实施差异化干预。在需求定义环节，采用价值映射矩阵识别潜在冲突维度；设计评审阶段嵌入可解释性约束检查清单，强制标注决策路径依赖节点；压力测试引入对抗性伦理场景集，覆盖公平性、鲁棒性与自主性三重边界；上线审计执行偏差阈值校验，对 $\mathcal{L}_{\text{fairness}}$ 与 $\mathcal{L}_{\text{privacy}}$ 进行量化回溯；运行监控则部署轻量级实时指标探针，动态捕获分布偏移与意图漂移信号。各节点均配套标准化工具包，确保原则内化具备可操作性、可验证性与可迭代性 [8]。

3.3 反馈驱动的原则再校准机制

反馈驱动的原则再校准机制构成动态化原则嵌入的闭环中枢，其本质是将伦理原则从规范性文本转化为可迭代的操作协议。该机制依托三类实证输入：偏差报告触发的场景归因分析、用户申诉映射的权利侵害图谱、第三方评估揭示的系统性缺口。每季度开展跨模态证据融合，通过权重分配算法对原则实施细则进行参数化修订，例如调整公平性约束中群体差异容忍阈值 ϵ_{group} 或重设透明度要求在高风险场景下的优先级系数 $\alpha_{\text{high-risk}}$。校准结果经多利益相关方审议后嵌入开发流水线，并同步更新合规检查清单与审计日志模板，确保原则演进可追溯、可验证、可执行 [9]。

4. 核心论点 B：制度设计的模块化------从统一规制到分层响应

4.1 风险等级划分的多维标尺

人工智能伦理风险的治理效能高度依赖于风险识别的精度与响应机制的适配性。本研究提出四维标尺构建动态风险光谱：影响范围界定技术应用的空间延展性与人群覆盖广度；不可逆程度评估系统失效后修复的可能性阈值；自主决策权重量化算法在关键环节中替代人类判断的深度；社会敏感性则捕捉公众认知、文化价值与制度信任的耦合强度。四维并非线性叠加，而呈现非对称交互效应------例如高社会敏感性叠加低不可逆性，可能触发强监管响应；反之，广域影响但可逆性强的场景，则倾向采用监测型治理。该多维标尺支撑模块化制度设计，使治理强度梯度匹配风险等级跃迁，避免统一规制导致的过载或失焦 [1, 10]。

4.2 治理工具箱的弹性组合策略

治理工具箱的弹性组合策略体现为风险等级与制度强度的精准适配。在高风险场景，强制性认证与第三方审计构成刚性约束机制，确保算法透明度与可追溯性；中风险场景则依托合规激励与结构化信息披露，通过正向引导提升组织自律水平；低风险场景聚焦能力建设与公众教育，强化社会主体的认知基础与参与能力。三类工具并非孤立运行，而是在统一风险评估框架下形成动态耦合：认证结果可转化为合规激励的准入依据，信息披露质量直接影响能力建设的针对性，公众反馈亦反向校准风险等级判定。该分层响应逻辑有效规避了"一刀切"规制导致的治理冗余或监管真空，使制度供给始终锚定技术演进的实际风险谱系，实现治理效能与创新活力的结构性平衡 [11]。

4.3 主体权责的动态再分配机制

人工智能系统生命周期中技术可控性呈现显著的动态衰减特征，其责任配置必须突破静态契约范式。开发者在模型训练阶段承担算法透明性与基础安全验证义务；当模型进入部署环节，部署者需对应用场景适配性、接口安全性及日志可审计性负主要责任；终端使用者则依据其专业能力与操作权限，在合理注意范围内履行操作合规义务；监管者职责随技术成熟度演进，从早期沙盒测试的规则制定者，逐步转向运行阶段的风险监测者与事后归责的裁决协调者。这种权责再分配并非线性转移，而是依据 $\mathcal{P}_{\text{controllability}}$ 阈值触发的条件响应机制，确保义务强度与实际控制能力严格匹配 [12]。

5. 深度对比与挑战：全球实践中的结构性差异与共性困境

5.1 区域路径差异：欧盟的规则先行、美国的市场驱动与亚洲的试验导向

欧盟强调规则先行，立法节奏快且具约束力，企业参与多限于合规响应；美国依托市场驱动，监管滞后于创新，企业深度嵌入标准制定与自我规制；亚洲国家则呈现试验导向特征，通过沙盒机制与政策试点快速迭代治理工具。三者在公众协商机制上差异显著：欧盟制度化程度高，美国依赖行业主导的多元协商，亚洲多采用有限范围的专家咨询与场景化听证。技术基础设施支撑能力亦呈梯度分布，欧盟强于监管科技部署，美国长于平台级治理工具开发，亚洲则聚焦垂直领域验证环境建设。结构性张力集中于立法刚性与技术演进速率的持续错配 [4]。

5.2 标准碎片化带来的合规成本与互操作障碍

全球人工智能伦理治理呈现显著的标准碎片化态势，多套并行的伦理框架、评估指标与认证程序在法域间缺乏协调机制。这种制度性重叠不仅大幅抬升企业的合规成本，更在技术验证、模型审计与跨境部署环节形成实质性互操作障碍。企业需重复适配不同监管要求下的风险评估流程与报告范式，导致资源错配与治理效能衰减。现有研究普遍认为，标准异构性已构成跨国技术协作的核心制度摩擦源，削弱了全球 AI 治理网络的整体韧性与响应敏捷性 [11]。

5.3 能力鸿沟与价值张力构成的深层制约

能力鸿沟在中小机构中表现为伦理评估工具缺失、专业人才匮乏与制度化流程缺位，导致风险识别滞后于技术部署节奏。多元文化语境下，价值排序呈现显著非线性冲突：隐私权重、透明度阈值与责任归属逻辑在不同法域间难以通约。更深层的张力源于短期商业目标与长期公共利益之间的结构性错配------算法迭代周期以月计，而社会影响评估需跨年度追踪，二者在时间尺度、绩效指标与问责机制上持续失谐。这种三重制约共同构成敏捷治理落地的底层瓶颈 [6, 11]。

6. 未来展望：构建兼具速度、韧性与正当性的中国路径

6.1 法治框架下的弹性授权机制

法治框架下的弹性授权机制要求基础性立法具备动态适配能力。在《人工智能法》等核心法律中，应系统性嵌入实施细则接口，明确授权主管部门依据技术演进节奏与风险图谱变化，发布分级分类管理目录及响应指南。该机制须以法定程序为边界，确保授权范围、启动条件与审查周期均具可预期性与可追溯性，避免行政裁量权无序扩张。同时，目录更新需配套影响评估与公众参与程序，保障制度韧性与正当性双重实现 [9]。

6.2 分层分类的风险响应体系

分层分类的风险响应体系需构建国家级风险预警平台、行业级合规服务中心与机构级伦理委员会三级架构，形成纵向贯通、横向协同的治理闭环。国家级平台聚焦宏观趋势研判与跨域风险识别，依托多源异构数据融合建模实现早期预警；行业级中心承担标准转化、能力建设与合规审计职能，推动通用原则向领域实践精准适配；机构级委员会则负责具体场景的伦理审查、动态评估与处置反馈，确保治理触达技术落地末梢。三者间须建立标准化接口与双向反馈机制，保障风险信息流、决策指令流与能力支持流的实时对齐与迭代优化 [2]。

6.3 面向公共价值的技术能力建设

面向公共价值的技术能力建设，须将伦理评估工具、偏见检测模块与影响模拟平台系统性嵌入国家人工智能基础设施体系。此举旨在实现合规能力的标准化供给与规模化复用，显著降低研发主体在数据采集、模型训练、部署验证等全链条中的伦理适配成本。通过统一接口规范、开放测试基准与可验证审计日志机制，支撑监管机构实施动态合规监测，同时赋能企业开展自主式伦理影响预判。技术能力的公共化配置，是平衡创新速率与治理深度的关键制度杠杆 [2, 11]。

7. 结论：迈向一种扎根实践、持续进化的治理智慧

7.1 敏捷治理的本质是制度的学习能力而非简化流程

敏捷治理的本质在于制度的学习能力，而非流程的简化或压缩。它要求治理体系内嵌结构化的反思机制、动态修正通道与知识沉淀规则，使制度本身具备持续适应性与进化韧性。这种能力体现为对新兴伦理风险的识别灵敏度、跨域协同的响应弹性以及治理经验的系统化转化效率，绝非临时性权宜之计或程序性让渡。

7.2 中国路径需平衡效率诉求与底线思维

中国路径的敏捷治理绝非效率优先的单向度演进，而须在技术创新激励与底线思维坚守之间构建动态张力。基本权利保障、国家安全与社会稳定构成不可让渡的刚性边界，任何敏捷化尝试若导致监管缺位或责任稀释，即构成治理异化。因此，制度设计需内嵌可回溯、可问责、可校准的约束机制，确保技术跃迁始终运行于法治化、伦理化与社会化协同的轨道之上。

7.3 治理现代化的终极指向是人本价值的稳定实现

敏捷治理的终极合法性不源于技术响应速度，而根植于人本价值的稳定实现。人的尊严、自主与福祉构成不可让渡的伦理基线，所有制度设计、流程迭代与工具部署均须以此为尺度进行持续校准。敏捷性仅是达成公共善的动态手段，而非目的本身；当治理实践偏离这一价值锚点，再高效的机制亦将丧失正当性根基。

参考文献：

[1] 王欧权. 人工智能在数字政府决策支持中的应用研究[J]. 电子通信与计算机科学, 2026, 8(4): 4-6.

[2] Ren B, Feng X. Ten Year construction achievements of Digital China and the advancement strategies for the 15th Five-Year Plan[J]. China Software, 2026, 41(5): 1-18.

[3] 张登科. 人工智能伦理风险与系统化治理研究[J]. 计算机与自主智能研究进展, 2026, 4(2): 106.

[4] 李昌奎, 张雅琪. 颠覆性技术研究（一）------理论起源、概念界定与五维分析框架[J]. 社会科学理论与实践, 2026, 8(1): 1.

[5] 聂佳磊. 数字经济背景下人工智能技术扩散的产业效应与发展管理研究[J]. 经济与管理发展研究, 2026, 2(5): 19-25.

[6] 肖泳衡, 俞洁, 周嘉淇. 从"资源垄断"到"普惠可及"：生成式人工智能对法律服务市场的赋能价值[J]. 人文学刊, 2026: 43-49.

[7] 张家媛. 生成式人工智能合规风险演化机制与动态治理范式------基于五维评估矩阵与法技协同的实证研究[J]. 社会科学进展, 2025, 14(6): 757-771.

[8] Roberts H, Cowls J, Hine E, Mazzi F, Tsamados A, Taddeo M, Floridi L. Achieving a 'Good AI Society': Comparing the Aims and Progress of the EU and the US[J]. Science and Engineering Ethics, 2021, 27(6): 68.

[9] Zhao B, Chen K. Transformation of the national innovation system under the artificial intelligence for science paradigm[J]. China Software, 2026, 41(5): 212-224.

[10] 欧权辉. 人工智能在数字政府决策支持中的应用研究[J]. 电子通信与计算机科学, 2026, 8(4): 4-6.

[11] 宋雁宾. 从伦理风险到韧性重构：人工智能治理的四维驱动机制与实践路径[J]. 现代管理, 2025, 15(7): 207-214.

[12] Callegaro M, Lozar Manfreda K, Vehovar V. Web Survey Methodology -- 网络调查方法[M]. 2026.