Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2025, 14(6), 1573-1583
Published Online November 2025 in Hans. https://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2025.146147

Hans汉斯

人工智能算法伦理风险的适应性治理研究

——基于浙江实践与欧美经验的整合框架

徐　丽[1]，徐　翌[2]

[1]浙江越秀外国语学院国际商学院，浙江　绍兴
[2]浙江越秀外国语学院数字贸易学院，浙江　绍兴

收稿日期：2025年10月23日；录用日期：2025年11月17日；发布日期：2025年11月26日

摘　要

人工智能算法伦理风险正成为制约技术赋能实体经济与包容性增长的关键瓶颈。本文以浙江省数字经济实践为研究场域，聚焦自动驾驶、AIGC等高频应用场景中的伦理治理挑战。基于“技术－社会－主体”三维分析框架，研究系统揭示了算法伦理风险的生成逻辑与多维呈现。研究发现，浙江在治理实践中虽展现出多元协同与敏捷治理特征，但仍面临治理科技鸿沟、标准缺失与协同梗阻等“系统性迟滞”。通过对欧盟刚性规制与美国柔性治理范式的比较与反思，本文构建了以监管沙盒为核心引擎的“基于风险的适应性治理”框架。该框架依托规制、技术与市场三大支柱，通过“风险识别－动态监管－效果评估－规则迭代”的闭环运行机制，实现技术创新与伦理规范的动态平衡。研究为解构算法伦理风险提供了系统理论视角，也为完善人工智能治理体系提供了兼具前瞻性与操作性的路径参考。

关键词

人工智能算法伦理，适应性治理，监管沙盒，浙江实践，欧美经验

Adaptive Governance of Ethical Risks in Artificial Intelligence Algorithms

—An Integrated Framework Based on Zhejiang Practice and European and American Experience

Li Xu[1], Yi Xu[2]

[1]School of International Business, Zhejiang Yuexiu University, Shaoxing Zhejiang
[2]School of Digital Commerce, Zhejiang Yuexiu University, Shaoxing Zhejiang

Received: October 23, 2025; accepted: November 17, 2025; published: November 26, 2025

文章引用: 徐丽, 徐翌. 人工智能算法伦理风险的适应性治理研究[J]. 人工智能与机器人研究, 2025, 14(6): 1573-1583.
DOI: 10.12677/airr.2025.146147

Abstract

Ethical risks in artificial intelligence algorithms are becoming a critical bottleneck that hinders technology from empowering the real economy and fostering inclusive growth. This paper takes the digital economy practices of Zhejiang Province as the research context, focusing on ethical governance challenges in high-frequency application scenarios such as autonomous driving and AI-generated content (AIGC). Based on a "technology-society-agent" three-dimensional analytical framework, the study systematically reveals the generative logic and multidimensional manifestations of algorithmic ethical risks. The research finds that although Zhejiang's governance practices exhibit characteristics of multi-stakeholder collaboration and agile governance, they still face issues of "systemic lag", such as gaps in governance technology, lack of standards, and coordination obstacles. Through a comparative analysis and reflection on the rigid regulatory paradigm of the European Union and the flexible governance approach of the United States, this paper constructs a "risk-based adaptive governance" framework with the regulatory sandbox as its core engine. Supported by three pillars—regulation, technology, and market—this framework achieves a dynamic balance between technological innovation and ethical norms through a closed-loop operational mechanism of "risk identification-dynamic supervision-effect evaluation-rule iteration". The study provides a systematic theoretical perspective for deconstructing algorithmic ethical risks and offers a forward-looking and actionable reference for improving the artificial intelligence governance system.

Keywords

AI Algorithm Ethics, Adaptive Governance, Regulatory Sandbox, Zhejiang Practice, European and American Experience

1. 问题的提出

人工智能正以前所未有的深度与广度重塑社会形态，其在驱动产业革新与提升社会效率的同时，也带来了责任主体模糊[1]、算法偏见歧视[2]、数据隐私泄露[3]等一系列复杂的伦理与治理挑战。这些挑战并非停留于理论推演，而是在技术落地的最前沿场景中集中爆发。在此背景下，如何构建一套敏捷、有效且与技术创新动态适配的伦理风险防控体系，已成为全球各国共同面临的紧迫课题。

中国作为人工智能发展的重要一极，其治理路径备受瞩目。伴随《新一代人工智能治理伦理规范》等国家顶层设计的推进，以及十五五期间“人工智能+”行动的深入实施，地方实践与制度探索的重要性日益凸显。浙江省凭借其深厚的数字经济底蕴与活跃的创新生态，孕育了杭州城市大脑交通调度算法、宁波智能制造生产优化算法、义乌跨境电商智能推荐算法等前沿应用，成为观察中国情境下算法伦理风险的“天然实验室”。但先发应用亦意味着先遇难题，在自动驾驶责任认定[1]、AIGC 知识产权争议[2]、金融算法歧视[3][4]等高频场景中，浙江省暴露出既有法律规范、监管架构与治理能力在应对新型风险的“系统性不适”。

[1]https://m.chinanews.com/wap/detail/zw/cj/2022/08-03/9818476.shtml
[2]http://www.legaldaily.com.cn/index_article/content/2025-08/08/content_9234742.html
[3]http://cq.people.com.cn/n2/2022/0815/c367643-40081456.html

与此同时，欧盟、美国等发达经济体在人工智能治理领域已率先进行制度性探索。欧盟通过《人工智能法案》确立“不可接受–高–有限–最低”四级风险监管框架，展现了严密的立法逻辑[5]；美国在人工智能治理方面采取了多维度的策略，包括立法、政策制定、行政命令、国际合作以及推动技术创新等[6]。这些经验为我们提供了宝贵参考，但任何制度的有效性都高度依赖于其赖以生存的社会土壤与制度环境。因此，本研究的核心任务并非简单移植欧美模式，而是致力于系统梳理浙江实践中的真问题与真需求，以欧美经验进行本土化整合与重构。基于上述考量，本研究将聚焦在如何基于浙江高频场景中的具体伦理风险，有机整合欧美治理经验中的有效成分，构建一个既立足中国地方实际又具备国际视野的算法伦理风险防控整合框架。

2. 人工智能算法伦理风险的多维透视与生成逻辑

算法伦理风险的表现具有显著的场景依赖性，其产生是技术内生缺陷、社会结构嵌入与主体责任缺失共同作用的结果，形成“技术–社会–主体”三维驱动的演化逻辑，这一框架得到指定文献的普遍认同[7]。

伦理风险的多维呈现

(1) 伦理风险类型

① 算法黑箱与可解释性缺失风险，即算法决策过程的“黑箱化”导致风险溯源与责任界定困难。ChatGPT 等大模型的生成内容虽逻辑连贯，但其参数运算与决策逻辑无法被人类完全解构，一旦输出错误信息，难以追溯问题根源[8]。进一步地，算法黑箱使得算法的运行逻辑和决策过程不透明，用户无法理解其背后的推理依据，这种不透明性削弱了公众对算法的信任[9]。

② 算法偏见与公平性受损风险，即算法通过训练数据继承并放大社会固有歧视，形成系统性偏见。由于地域、种族、年龄、性别、收入、教育等背景的不同，人类社会中始终存在着各种偏见或歧视，并有可能在人工智能算法的推动下迅速扩大或逐步加深[10]。教育人工智能中的算法偏见已常态化，主要集中于性别、种族与地域差异等维度[11]；算法偏见列为数字金融“四大新型风险”之首，其直接侵犯金融消费者的公平权，模型误读把统计相关性当作因果性，例如以“某地区违约率高”为由对整个地区收紧信贷，造成结构性不平等[12]。

③ 数据滥用与隐私侵犯风险。算法对海量数据的依赖导致隐私保护边界模糊。隐私侵犯风险发生可能性，远高于偏见歧视与权责不清，如服务机器人为了提升服务体验，过度采集用户表情、动作、声纹等生物特征信息，且常默认勾选“同意”，用户并不知情[3]；同样，AIGC 训练存在未经授权使用用户数据的风险，隐私合规性成为核心伦理隐患，亟需强化授权验证与伦理审查[13]。

④ 责任归属与问责困境风险。算法自主性增强使责任链条断裂，难以界定过错主体。人工智能自主发明的专利归属争议尚无明确法律依据，算法开发者、训练数据提供者与模型使用者的责任划分陷入僵局[1]；有学者研究发现以 Sora 为例，其生成虚假视频引发的名誉侵权事件中，平台与开发者相互推诿，导致受害者维权无门[14]。

不同领域人工智能应用的主要风险与治理趋势可归纳如下(见表 1)。

Table 1. Main risks and governance trends of artificial intelligence applications in different fields
表 1. 不同领域人工智能应用主要风险和治理趋势

领域	核心技术	协同机制	主要风险	治理趋势
金融[4]	大数据/算法/自动化决策	准入备案/实时监测/ 事后评估	数据安全、黑箱问题	伦理入法、治理前移、多元共治
自动驾驶[15]	多模态感知、端到端决策模型	V2X 协同控制、异构系统耦合	公平、透明、归责	数字孪生验证、AI 安全可解释、监管沙盒

续表

教育[11]	算法检测工具、可解释性 AI 技术	多元主体协作、技术–人文融合	数据隐私、教育公平	多场景伦理研究、检测工具开发
医疗[16]	机器学习、数据挖掘、自然语言处理、图像处理	多主体共治、全周期监管、跨部门协同	算法偏见、隐私泄露、黑箱问责、实验次生风险	全流程伦理治理、动态风险防控、技术伦理融合

数据来源：根据文献整理。

(2) 风险生成的深层逻辑

① 技术内生维度存在算法与数据的双重缺陷。技术层面的固有特性构成风险产生的物质基础。算法设计方面，强化学习的“奖励机制偏差”导致模型优先追求效率而忽视伦理，如推荐算法为提升点击量推送低俗内容，服务机器人为降低能耗简化安全校验[17]；大模型参数规模扩大引发的“涌现现象”增加行为不可预测性，Sora 生成视频时出现“物体悬浮”等物理逻辑错误，便是典型表现[14]。数据处理人员的价值观、个人偏见会影响数据结构，导致数据质量不高，进而使训练出的算法先天缺乏可靠性，甚至存在算法偏见[9]；研究显示，教育大数据自带历史偏差、采样不充分等问题，为算法风险埋下源头隐患。算法的思维处理惯性、黑箱特性及反偏见设计缺失，会放大数据偏差，最终诱发算法偏见甚至歧视[11]。

② 社会交互维度凸显技术理性与制度滞后的失衡。技术在社会交互中的快速发展与制度规范的滞后形成矛盾，导致决策自主受控、隐私侵犯等伦理风险频发，难以有效约束技术应用[3]；在教育领域的研究印证了这一观点——例如自适应学习平台主导学习路径推荐、智能答疑机器人替代部分师生沟通，使交互更重效率却弱化了情感共鸣与人文关怀，针对这方面的规范仍然不够[18]。制度层面，传统“风险基准”监管模式无法适配算法快速迭代，生成式 AI 出现后，现有法规对“虚假内容生成”“算法歧视”的规制存在真空[19]。

③ 主体责任意识与风险感知的缺失。利益相关者的认知局限与责任缺位加剧了伦理风险。从从业者层面看，技术主体的责任意识薄弱，在设计与决策中易因价值偏差与利益追求而忽视长远伦理责任；同时，人类对算法黑箱及系统复杂性的风险认知本身存在局限，导致对其潜在伦理后果的预见与评估能力不足[20]；OpenAI 的数据泄露事件、医疗 AI 企业未校验训练数据偏差导致误诊，均凸显企业责任意识薄弱。在对于 ChatGpt 使用分分析中，设计者责任缺失，即将其偏见嵌入算法，缺乏伦理约束；使用者责任与风险感知缺失，即滥用技术，忽视其对自身和社会的危害；人类整体风险感知局限，受认知所限，难以预见和评估复杂风险，导致主体责任意识集体弱化[21]。

3. 浙江实践：治理探索、深化与典型困境

作为中国数字经济的先行区与制度创新的“试验田”，浙江省在人工智能算法伦理风险治理中展现出鲜明的系统性制度效能。其治理实践超越了单一工具或领域的局限，形成了“边发展、边治理、边完善”的动态响应机制，构建了多维度、成体系的综合治理框架。然而，该机制的治理效能亦受限于若干结构性与本源性的困境。

3.1. 政策演进：从产业促进到伦理规制的体系化构建

浙江省的算法治理政策体系，经历了从聚焦经济产业到统筹发展与安全的演进过程，形成了层次分明、相互支撑的政策框架。

第一阶段(至 2023 年)：以产业发展与基础设施构建为核心。此阶段的政策重心在于为人工智能产业奠定基础、营造环境。以《浙江省促进新一代人工智能发展行动计划(2019~2022 年)》等文件为代表，政策目标集中于技术研发、产业集聚和应用场景开拓。此时，算法主要被视为提升效率的工具，其伦理与

社会风险尚未成为独立的政策议题。

第二阶段(2024 年至今)伴随生成式人工智能的爆发式应用，浙江省的政策制定呈现出明显的“规制转向”，形成了“省级宏观指导－行业领域规范－地方创新实践”的多层次治理体系。在宏观层面，《浙江省人民政府关于推动数字经济创新提质发展的实施意见》中明确提出要“探索建立人工智能伦理规范和治理框架”。在行业领域，2025 年密集出台的《浙江省推进“人工智能 + 教育”行动方案(2025~2029年)》《浙江省人工智能产业高质量发展实施方案》等文件均专设章节强调“安全可控和伦理治理”。

3.2. 实践特征：多元协同与敏捷治理的初步显现

浙江的治理实践，超越了传统的政府单一监管模式，在多个前沿领域展现出以下三个鲜明特征：

一是构建“多元共治”的协同生态。浙江初步形成了“政府引导、行业自律、企业主体、社会监督”的治理生态。政府角色从“划桨者”向“掌舵者”转变，通过制定规则、划定红线进行引导。行业协会(如浙江省人工智能产业联盟)积极制定行业伦理公约，搭建交流平台。头部企业(如阿里巴巴、海康威视)率先建立内部算法伦理委员会，开展自我规制。此外，通过开通举报渠道、发布典型案例等方式，社会监督力量被初步激活，一个多方参与的治理网络正在形成。

二是践行“包容审慎”的监管创新。“包容审慎”是浙江处理创新与监管关系的基本准则。其核心在于对新技术、新业态设置一定的“观察期”与“弹性空间”，而非一上来就“管死”。在自动驾驶领域，这一原则得到了生动体现。浙江省及杭州市已开放多个智能网联汽车测试区域[4]，允许企业在真实道路环境中测试其自动驾驶算法。这种“监管沙盒”式的管理，为算法迭代提供了宝贵的真实数据，同时要求企业建立完善的数据记录、安全预警和远程接管机制，将风险控制在可控范围内。这正是在“包容”与“审慎”之间寻求平衡的典型实践。

三是形成“场景驱动”的治理路径。浙江的算法治理强调与具体产业和社会治理场景深度融合，实现精准施策。在智慧医疗领域，浙江通过建设健康医疗大数据中心，构建了数据权限管控、脱敏、加密、水印等安全基础能力，创新“数据可用不可见”的授权运营机制(如安全沙箱、多方可信环境)，并明确原始数据不出域、数据产品合规流转的原则[5]。在金融科技领域，地方金融监管局关注信贷算法的公平性问题，防范算法偏见导致的“数字红线”[6]。在自动驾驶领域，治理焦点则集中于算法决策的安全性与事故责任界定。例如，在宁波市的智能网联汽车测试规程中，已开始探索要求企业对自动驾驶系统在特定场景(如“碰撞不可避免”时)的决策逻辑进行说明和备案，这为未来厘清算法责任迈出了重要一步。在智慧教育领域，杭州的实践则从源头培育伦理意识。这种“场景化”治理，使抽象的伦理原则在具体业务逻辑中找到落脚点，提升了治理的针对性与有效性。

3.3. 现实困境与典型场景剖析

浙江虽构建了较为完善的政策体系，但在从“文本治理”迈向“行动治理”时，仍面临诸多挑战。一是“治理科技”鸿沟，中小企业因专业算法审计与偏见检测工具成本高昂且操作复杂，合规难度大；二是标准体系缺失，现有政策多为原则性规定，缺乏细化的行业标准和可操作的合规指南，致使企业合规成本高，监管执法尺度不一；三是协同治理与人才支撑短板，算法治理涉及多部门，权责边界与协同机制尚不清晰，复合型治理人才稀缺，成为制约治理体系现代化的关键瓶颈。

在典型应用场景中，这些挑战表现得尤为突出。浙江自动驾驶落地实践中，多起真实案例为算法伦

[4]https://www.hangzhou.gov.cn/art/2025/3/12/art_812266_59110262.html
[5]https://wsjkw.zj.gov.cn/art/2024/7/31/art_1229129056_5340402.html
[6]https://zjic.zj.gov.cn/zkfw/yjzx/202401/t20240108_21444072.shtml

理治理提供具象样本，集中折射核心争议。2025 年杭州临平区案件 [7]中，车主王某某(血液乙醇含量 114.5 mg/100 ml)加装“智驾神器”规避 L2 级辅助驾驶监测，使车辆主驾无人行驶 20 分钟，最终以危险驾驶罪获刑，凸显公众对“人机共驾”伦理边界的认知偏差及技术工具异化风险。2024 年 12 月杭州余杭区事故更典型，车主沈先生驾车出车位时与合法路权的新石器无人快递车碰撞，数据显示无人车减速至 6 km/h 后突加速至 25 km/h；虽交警初判沈先生主责，但公众与当事人对“无人系统应具安全冗余”的诉求，及事故暴露的算法动态障碍物响应异常，直指智能驾驶“安全优先”伦理原则的实践落差，也反映法律框架对人机混合责任划分的适配难题。上述案例共同揭示浙江自动驾驶推广中，技术伦理设计、用户责任认知与法律责任体系的现实张力。

浙江 AIGC 落地实践中，多起真实案例集中暴露知识产权与虚假信息治理双重难题。杭州某平台利用大模型生成“网红餐厅”评测，因模型学习不实信息错误标注“卫生不达标”，导致商家商誉受损，平台以“AI 自主生成”规避审核责任，这与杭州中院 2025 年判决的 AI“伪种草”[8]文案案核心争议相通，均指向技术中立抗辩无法豁免平台注意义务的司法认定。上述案例印证了 AIGC 领域知识产权归属模糊与虚假信息溯源困难[14]的现实，暴露出现有法规在界定生成内容侵权边界与平台审核责任上的滞后性。

4. 欧美经验：两种范式的比较与反思

在全球人工智能治理的谱系中，欧盟与美国分别代表了两种风格迥异却极具代表性的规制路径。本章旨在超越对欧美政策文本的简单描述，深入剖析其演进逻辑、内在张力与最新动态，特别是通过对欧盟撤回《人工智能责任指令》这一关键事件的解读，揭示其对于浙江乃至中国治理实践的深刻启示。

4.1. 欧盟路径：从统一立法到审慎回调的合规驱动模式

欧盟始终致力于构建一个统一、刚性且以基本权利保护为核心的人工智能治理框架，其核心是“基于风险”的合规驱动模式。欧盟《人工智能法案》的通过，标志着全球首部全面监管 AI 的综合性法律诞生[22]。其核心在于依据风险等级对 AI 系统进行“金字塔式”的分类监管，其具体风险分类与典型应用场景如表 2 所示。这一体系旨在通过清晰的法律义务，为市场建立稳定预期，但其复杂性也给企业，尤其是中小企业，带来了沉重的合规负担。

Table 2. Risk classification and corresponding typical application scenarios under EU legislation
表 2. 欧盟法案风险分类与典型应用场景

风险等级[5]	核心治理原则	典型应用场景
不可接受的风险	全面禁止使用	1. 社会评分系统(基于多维度数据对个人社会信用、行为进行量化评级并关联公共服务权益)、 2. AI 自主武器系统(具备自主目标识别与攻击能力的致命性武器)
高风险	严格事前准入与全程管控	3. 关键基础设施 AI (如能源电网调度系统、城市交通信号控制中枢) 4. 医疗 AI 设备(如手术机器人、辅助诊断系统) 5. 自动驾驶系统(L2 +驾驶辅助、L3~L4 自动驾驶)
有限风险	以透明度为核心	6. 聊天机器人(如客服 AI、对话式交互工具) 7. 内容推荐算法(如电商商品推荐、资讯推送系统) 8. 虚拟助手(如语音交互类智能终端)
最小风险	原则自由使用，鼓励自律	9. 个人 productivity 工具(如 AI 文档校对、简单数据统计工具)

数据来源：根据公开资料整理。

[7]https://sifa.wenzhou.gov.cn/col/col1680223/art/2025/art_5d477525d1274771bb7ef15b91ec6d7b.html
[8]https://www.hzzx.gov.cn/cshz/content/2025-09/09/content_9079192.htm

2025 年初欧盟委员会正式撤回《人工智能责任指令》提案，是理解欧盟治理困境的一个分水岭事件。该指令草案旨在通过“因果关系推定”和“证据披露请求权”来减轻原告的举证负担。然而，其撤回深刻暴露了刚性立法在技术现实面前的局限性，一是技术复杂性，在算法的“黑箱”面前，即使实行举证责任倒置，原告方往往仍缺乏足够的技术知识来启动诉讼；二是创新抑制担忧，过于宽松的责任规则将极大地抑制投资与创新；三是与《AI 法案》的重叠与冲突。这一撤回标志着欧盟治理思想从“立法万能”向“务实审慎”的转变。

4.2. 美国路径：在行业自律与分散立法间的协同治理模式

与欧盟的顶层设计不同，美国的 AI 治理呈现出“轻触式”联邦监管与“碎片化”州级立法并存，并高度依赖行业自律与市场机制的特点。在联邦层面，主要通过白宫发布的《人工智能权利法案蓝图》等非强制性文件确立治理原则。在立法上则采取“问题导向”的分散策略。这种路径避免了全面的监管框架，赋予行业更大的灵活性。在市场层面，人工智能责任保险作为一种风险转移工具正在快速发展，保险公司通过核保流程实质上扮演了“私人监管者”的角色，推动企业提升风险管理水平。同时，通过集体诉讼和司法判例，普通法体系也在逐步厘清算法应用中的责任边界。

对欧美路径的比较分析揭示，试图以一部法律细致规制所有人工智能场景具有固有局限性。这一发现为浙江的治理实践提供了超越简单模仿的深刻启示，即一个有效的治理框架必须内置制度弹性。这一认知具体体现于规制设计的差异化，要求在守住安全底线的前提下，为不同风险等级与发展阶段的算法应用设计精准的合规路径，例如通过为中小企业提供简化的合规指南或豁免条款，避免“一刀切”带来的创新抑制。在此基础之上，需着力填补治理科技的鸿沟，鉴于欧盟合规评估与美国算法审计的有效性均深度依赖可靠的技术工具，浙江必须将发展本土治理科技产业与建设公共检测平台提升至战略高度，从而系统性降低合规与监管的双重成本。最终，治理结构的优化关键在于构建多元协同的闭环生态，美国的经验表明市场机制与司法救济是行政监管不可或缺的补充。这意味着浙江需积极鼓励算法责任保险、第三方审计等市场化工具体系的发展，同时完善司法救济渠道并大力推动行业自律，最终形成政府、市场与社会协同发力的良性治理格局。

5. 整合框架构建：“基于风险的适应性治理”框架

在系统剖析浙江实践困境与深度反思欧美经验模式的基础上，本章旨在提出一个扎根中国情境、融通国际视野的整合性治理框架。该框架以监管沙盒为核心实施载体，致力于在鼓励创新与防控风险之间构建一种动态的、可持续的平衡。

5.1. 核心原则：框架的三大基石

本框架的构建立足于三大核心原则，它们共同决定了治理体系的基本方向与特性。

① 基于风险角度承认“一刀切”治理的低效与不切实际，转而依据人工智能算法的应用场景、潜在影响范围与可能危害程度进行风险等级划分。对高风险应用(如自动驾驶、疾病诊断)实施严格监管，对低风险应用(如娱乐推荐)采取柔性引导，从而实现监管资源的优化配置。

② 敏捷灵活是治理体系必须具备学习与演进的能力。这意味着监管规则不应是僵化的，而应能根据技术迭代、市场反馈和风险评估的变化进行动态调整。监管沙盒是本原则的核心体现，为创新提供安全的测试空间，并为规则迭代提供现实依据。

③ 多元协同下明确政府、企业、行业组织、用户与公众等多元主体的权责利，构建“赋能式治理”生态。政府的角色从“全能监管者”转变为“沙盒环境的构建者与治理底线的守护者”。

5.2. 三大支柱：框架的支撑体系

为实现上述原则，本框架由三个相互支撑的支柱构成，覆盖规制、技术与市场三个维度。

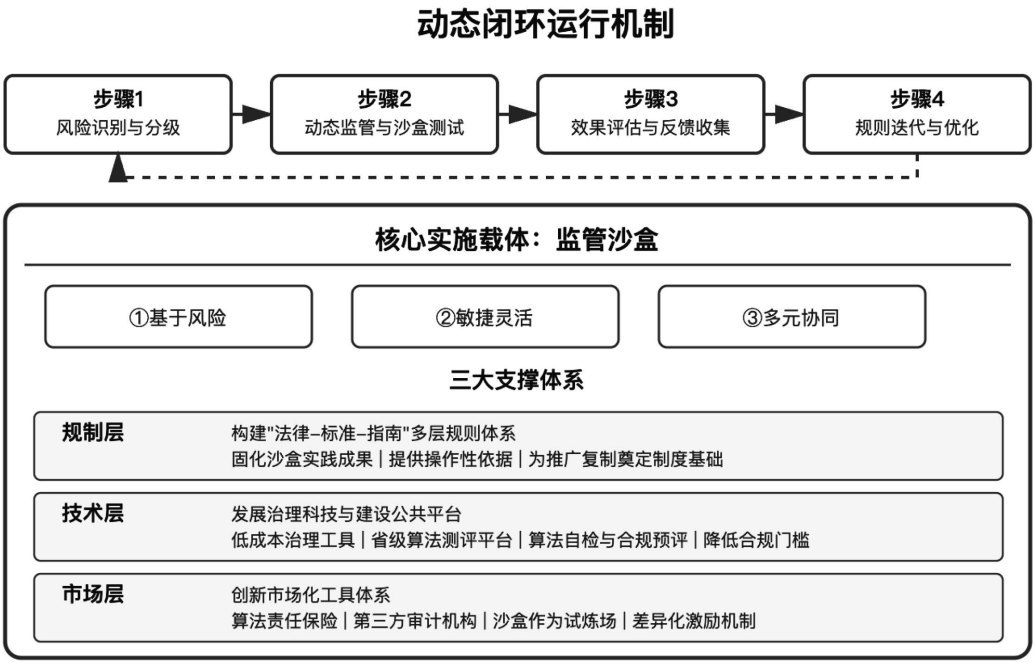
支柱一为规制层——构建“法律–标准–指南”多层规则体系。旨在解决“原则性规范与操作性指南断层”问题。其核心任务是将沙盒内已验证有效的治理措施，及时固化为地方标准、团体标准与行业合规指南，为企业提供清晰、可测量的技术依据，并为沙盒的推广复制奠定制度基础。

支柱二为技术层——发展治理科技与建设公共平台。旨在弥合“治理科技鸿沟”。通过政策扶持研发低成本治理科技工具，并建设省级算法测评公共服务平台。该平台应深度服务于沙盒生态，为入盒企业，尤其是中小企业，提供普惠性的算法自检、合规预评与技术支援，大幅降低合规门槛[23]。

支柱三为市场层——创新市场化工具体系。旨在引入市场化力量，形成对行政监管的有效补充。关键是鼓励开发“算法责任保险”并扶持独立的第三方算法审计机构。沙盒是这些市场工具最理想的“试炼场”和“首发站”，通过差异化保费和权威审计报告，激励企业主动提升治理水平。

5.3. 运行机制：框架的动态闭环

三大支柱的有效运转，依赖于“风险识别–动态监管–效果评估–规则迭代”的闭环运行机制，而监管沙盒正是驱动这一闭环的核心引擎。一是风险识别与分级，建立动态评估机制，为沙盒的准入与分级监管提供依据；二是动态监管与沙盒测试，根据风险分级，对高风险系统强制审计，对创新性应用则准入“监管沙盒”，在真实场景中有限测试，并配套保险等风险缓释工具；三是效果评估与反馈收集，通过公共服务平台、保险数据、审计报告等渠道，重点收集沙盒内治理措施的执行效果信息；四是规则迭代与优化，基于沙盒实践的反馈，定期审视和修订标准、指南乃至法规，完成从实践到规则再到实践的闭环学习，确保算法伦理治理框架具备持续的适应性与活力。该“基于风险的适应性治理”框架的整体结构与运行流程如图 1 所示。



数据来源：根据文字整理绘制。

Figure 1. Risk-based adaptive governance framework
图 1. 基于风险的适应性治理框架

6. 结论与建议

6.1. 结论

通过整合浙江实践与欧美经验，本研究核心结论如下：其一，人工智能算法伦理风险本质是技术不确定性与社会结构性风险的交织，需超越技术修复，构建系统性制度设计；其二，单一刚性立法监管存在局限(如欧盟《人工智能责任指令》撤回)，治理需“刚柔并济”以平衡创新与规范[5]；其三，浙江治理探索虽领先，但面临“治理科技落差”“合规标准断层”“多元协同梗阻”，需规制、技术、市场三轮驱动破局；其四，研究提出的“基于风险的适应性治理”框架，依托规制、技术、市场三大支柱及“风险识别－动态监管－效果评估－规则迭代”闭环，将“监管沙盒”升维为核心实施载体，为解决困境提供兼具理论自洽性与实践操作性的方案。

6.2. 建议

(1) 浙江省层面：深化沙盒工具应用，精准化解伦理风险

在省级层面，深化监管沙盒的应用是精准化解算法伦理风险的核心路径。具体而言，建议从以下三个维度协同推进：

首先，推动沙盒从政策试验场向规则生成器演进。实现这一目标，需由省政府部门牵头，联合之江实验室、省标准化研究院等科研机构，依托产业沙盒的实践基础，优先针对自动驾驶、AIGC、智慧金融等伦理风险高发领域，提炼沙盒内已验证有效的伦理防控措施，明确算法公平性、透明度、责任划分的具体指标，形成明确算法公平性、透明度与责任划分的地方性技术标准，为企业提供“风险可测、合规有据”的操作范式，并为国家立法供给经过实证检验的规则样本。

其次，构建基于能力分层的协同治理网络，以破解中小企业伦理审查能力不足的瓶颈。依托沙盒生态，借鉴宁波沙盒政企协同模式，开发适配沙盒场景的轻量化伦理自检工具；参考杭州沙盒数据流通合规经验，提供标准化伦理评估数据集，并通过“科技创新券”补贴入盒企业的第三方伦理审计费用，系统性降低中小企业伦理合规成本。

最后，强化沙盒激励约束，培育伦理友好型内生生态。扩大杭州、宁波算法伦理风险防控专项沙盒的准入范围，将企业在沙盒内的伦理表现(如算法偏见整改成效、用户隐私保护力度)与专项金融支持深度挂钩；借鉴欧美“算法责任保险”实践，联合保险机构为沙盒内企业开发定制化伦理风险保险产品，覆盖因算法伦理缺陷引发的侵权赔偿风险，通过“激励 + 保障”双轮驱动，引导企业主动防控伦理风险。

(2) 国家层面：从地方实践到国家治理体系的制度集成

在国家层面，构建人工智能伦理治理体系的战略任务，在于将浙江沙盒的先行经验进行系统性的制度集成与能力推广。其顶层设计应围绕以下三个关键方向展开：

其一，推动实践成果的法治化转型，为全国治理提供规范性基础。推广浙江沙盒经验，完善伦理治理法律框架建议在《人工智能法》立法进程中，吸纳浙江沙盒在伦理风险责任划分上的实践成果(如沙盒“开发者－部署者－使用者”三级责任追溯机制)，专设条款明确自动驾驶、AIGC 等场景的伦理责任边界；参考浙江沙盒对“算法黑箱”的应对逻辑，设立“伦理缺陷过错推定”规则，为地方监管与司法实践提供上位法依据，同时将浙江列为“国家算法伦理治理沙盒示范区”，允许其在伦理审查机制、风险补偿制度上进一步探索。

其二，构建开放协同的治理知识系统，实现国际经验的本土化融合。以沙盒为纽带，推动欧美经验本土化落地借鉴欧盟《人工智能法案》中“高风险 AI 系统”分级监管思路，结合浙江沙盒的产业适配经验，构建“国家－省－市”三级沙盒体系，在国家级沙盒中试点欧美成熟的伦理工具(如美国 XAI 计划的

可解释性技术、欧盟的伦理影响评估流程)，并通过浙江沙盒的实践验证其本土化适配性；鼓励浙江沙盒与欧美重点实验室、企业建立伦理风险防控协作机制，联合开展算法公平性、隐私保护等技术攻关，推动国际经验与本土实践的深度融合。

其三，启动面向治理现代化的人才培养计划，强化国家层面的能力储备。结合杭州、宁波沙盒对“算法技术 + 伦理法律 + 行业场景”人才的需求，在高等教育体系中增设“算法伦理治理”交叉学科方向，支持高校与浙江沙盒企业共建实践基地，定向培养具备沙盒伦理风险研判能力的高层次人才。针对监管队伍，则应实施以浙江沙盒审查流程与风险案例为核心的“监管能力现代化”专项培训，提升全国监管队伍的伦理风险防控能力。

基金项目

本文系 2024 年度浙江省软科学研究计划项目《浙江人工智能应用的科技伦理风险评估及其治理路径研究》(编号 2024C35048)研究成果之一。

参考文献

[1] 刘鑫. 人工智能自主发明的伦理挑战与治理对策[J]. 大连理工大学学报(社会科学版), 2023, 44(4): 80-85.

[2] Ghasemaghaei, M. and Kordzadeh, N. (2024) Ethics in the Age of Algorithms: Unravelling the Impact of Algorithmic Unfairness on Data Analytics Recommendation Acceptance. *Information Systems Journal*, **35**, 1166-1197. https://doi.org/10.1111/isj.12572

[3] 李梦薇, 徐峰, 晏奇, 等. 服务机器人领域人工智能伦理风险评估方法的设计与实践[J]. 中国科技论坛, 2023(10): 74-84.

[4] 江军, 李牧翰. 人工智能金融领域应用伦理风险及其法律治理[J]. 江西财经大学学报, 2025(1): 127-136.

[5] 谢波, 陈晨. 人工智能算法安全治理的欧美经验与中国路径[J]. 中国科技论坛, 2025(8): 172-180.

[6] 陈龙, 刘刚, 戚聿东, 等. 人工智能技术革命: 演进、影响和应对[J]. 国际经济评论, 2024(3): 9-51+4.

[7] 黄静秋, 邓伯军. 人工智能算法的伦理规制研究[J]. 北京科技大学学报(社会科学版), 2025, 41(2): 88-96.

[8] 李子浩, 李天云. 技术向善: 人工智能大模型伦理风险识别及治理路径[J]. 学术交流, 2024(8): 30-42.

[9] 陈雄燊. 人工智能伦理风险及其治理——基于算法审计制度的路径[J]. 自然辩证法研究, 2023, 39(10): 138-141.

[10] 赵志耘, 徐峰, 高芳, 等. 关于人工智能伦理风险的若干认识[J]. 中国软科学, 2021(6): 1-12.

[11] 王佑镁, 王旦, 王海洁, 等. 算法公平: 教育人工智能算法偏见的逻辑与治理[J]. 开放教育研究, 2023, 29(5): 37-46.

[12] 杨松, 周楠. 数字金融的算法风险及其法律规制[J]. 陕西师范大学学报(哲学社会科学版), 2024, 53(2): 40-54.

[13] 郑煌杰. 生成式人工智能的伦理风险与可信治理路径研究[J]. 科技进步与对策, 2025, 42(12): 38-48.

[14] 刘祖兵. 论 Sora 的伦理风险与我国治理因应——也谈我国参与人工智能算法伦理全球治理的基本路径[J]. 河海大学学报(哲学社会科学版), 2024, 26(5): 99-112.

[15] 曹建峰. 论自动驾驶汽车的算法安全规制[J]. 华东政法大学学报, 2023, 26(2): 22-33.

[16] 于雪, 刘博涵. 医疗人工智能社会实验的伦理风险及其治理路径[J]. 医学与哲学, 2025, 46(4): 45-50.

[17] 张宁, 高鹏程. 生成式人工智能情感模拟的伦理风险与治理路径: 基于技术-社会互构理论框架的分析[J]. 科学决策, 2025(2): 123-134.

[18] 王斌伟, 付圣莹. 人工智能教育应用中的伦理风险及其应对[J]. 学术研究, 2025(4): 1-9.

[19] 许建峰, 吕昭诗. 人工智能分类分级监管研究——从风险基准到重要性导向的逻辑转变[J]. 济南大学学报(社会科学版), 2025, 35(5): 148-160.

[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] 邹开亮, 刘祖兵. ChatGPT 的伦理风险与中国因应制度安排[J]. 海南大学学报(人文社会科学版), 2023, 41(4): 74-84.

[22] (2025) Artificial Intelligence Index Report 2025. https://hai-production.s3.amazonaws.com/files/hai-ai-index-2025-policy-highlights.pdf

[23] 董克, 宋雨宸, 吴佳纯. 欧盟人工智能数据治理的政策布局与治理特征研究[J]. 农业图书情报学报, 2025, 37(7): 4-18.