实验主义治理：人工智能创新发展与治理模式革新

Experimentalist Governance:

The Development of Artificial Intelligence Innovation and the Reform of Governance Modes

徐磊 /XU Lei

（南京农业大学人文与社会发展学院，江苏南京，210095）
(College of Humanities and Social Development, Nanjing Agricultural University, Nanjing, Jiangsu, 210095)

摘　要：当前，人工智能技术已经深度融入经济社会各个行业领域，重塑人类生产生活模式，促进生产力革命性跃迁和生产关系深层次变革。实验主义治理作为应对复杂性与不确定性的新型治理模式，以实验为核心逻辑，具备治理主体多元协同、治理依据灵活动态、治理过程积累迭代、治理结果创新可控等特点。为了有效平衡人工智能创新发展与风险规制，推动人工智能技术朝着增进人类共同福祉的方向健康有序发展，我国有必要构建人工智能实验主义治理体系，确立创新可续、风险可控、价值可塑的三位一体治理目标体系，完善政府统筹、企业主导、科研支撑、公众参与的多元协同治理体制，健全风险管理、信息沟通、权利保障、责任豁免等治理机制。

关键词：人工智能　实验主义治理　复杂适应系统　以人为本　风险管理机制

Abstract: Artificial intelligence has been deeply integrated into various industries and realms of the economy and society, reshaping the mode of production and life and driving a leap in productivity as well as a deep change in production relations. As a new governance paradigm for addressing complexity and uncertainty, experimentalist governance takes experimentation as its core logic and possesses characteristics including multi-stakeholder collaboration, flexible and dynamic governance bases, cumulative and iterative governance processes, and controlled but innovative governance outcomes. To effectively balance technological innovation and risk mitigation, and promote the healthy and orderly development of AI in the direction of advancing the common well-being of humanity, China needs to establish an experimentalist governance system for AI. This involves setting up a tripartite system of governance featuring goals of sustainable innovation, controllable risks, and malleable values; improving a multi-stakeholder collaborative governance structure with government coordination, enterprise leadership, scientific research support, and public participation; and refining governance mechanisms of risk management, information communication, rights protection, and liability exemption.

Key Words: Artificial intelligence; Experimentalist governance; Complex adaptive systems; Human-centered approach; Risk management mechanisms

中图分类号：G206.2; TP18　DOI: 10.15994/j.1000–0763.2026.08.012　CSTR: 32281.14.jdn.2026.08.012

基金项目：江苏省社会科学基金后期资助项目“生成式人工智能安全风险的法律治理研究”（项目编号：23HQB029）。
收稿日期：2025 年 12 月 22 日
作者简介：徐　磊（1986–）男，河北衡水人，南京农业大学人文与社会发展学院副教授，研究方向为科技伦理。Email: xulei@njau.edu.cn

一、实验主义治理何以兴起：背景与共识

随着模型算法的持续优化、计算能力的显著跃升以及数据规模的爆发式增长，人工智能技术飞速发展，在医疗、金融、教育、交通等诸多领域展现出强大的创新能力和变革力量。人工智能系统在不断突破人类传统认知的同时，也引发人工智能创新发展与既有治理模式之间的深层矛盾。基于法律的治理以明确行为规范、设置权利义务为核心，依托法律的制定、执行、监督等闭环结构，在网络安全、数据安全、个人信息保护等领域发挥着重要的作用。然而，人工智能系统所具有的创新性、复杂性、智能性等，与法律治理所体现的稳定性、规范性、确定性之间存在结构性张力。这使得基于法律的治理在人工智能创新发展中存在诸多不足，难以适配人工智能系统治理的现实需求。基于设计的治理主张，在人工智能系统的设计阶段就需要充分考虑法律法规、价值观念、利益诉求等各种治理要求，并且将其转化为设计特征和具体功能，从而使人工智能系统能够遵循相关治理要求。[1]

可是，基于设计的治理难以跟上人工智能系统创新发展的步伐，并且不同设计部门可能根据自身职责和目标制定不同的设计策略，使得基于设计的治理缺乏协调性和一致性。基于风险的治理是指通过科学的方法识别、评估、应对人类社会和自然世界中的潜在风险，以此将社会、经济、环境等方面的负面影响最小化，同时最大程度地保障公共利益和社会安全。[2]不过，生成式人工智能和未来可能实现的通用人工智能以其多模态、跨领域、智能化特性，打破了既有风险分级的线性逻辑，由此暴露出风险分级方法的深层缺陷。[3]人工智能创新发展所伴随的衍生风险可能在多年后才会完全显现，这使得既有治理模式难以对其进行有效识别与管控。

实验主义治理是应对现代社会复杂性与不确定性的新型治理模式。实验主义治理的核心包括“框架目标与衡量指标、结合实际与自主探索、成效评估与对标分析、调整优化与循环往复”。[4]“实验主义治理始于实验设计，包括确定实验目的、选择实验对象和方法等；实验过程中产生大量数据和经验，这成为知识生产的基础；实验主义治理的最终目标是将实验成果转化为政策，推动政策变革。”[5]实验主义治理已经应用于金融科技领域、[6]自动驾驶领域、[7]生物工程食品领域等。[8]当前，学术界对于实验主义治理的研究呈现出三大趋势：第一，从治理工具转向治理活动，强调实验设计与权力结构的互动；第二，从单一领域治理扩展到跨域综合治理，关注地方实验与全球规则的衔接；第三，从描述性案例研究转向规范性理论构建，试图建立实验主义治理的制度体系。

习近平总书记指出：“面对新一代人工智能技术快速演进的新形势，要充分发挥新型举国体制优势，坚持自立自强，突出应用导向，推动我国人工智能朝着有益、安全、公平方向健康有序发展。”[9]《国务院关于深入实施“人工智能+”行动的意见》明确指出：“深入开展人工智能社会实验。”[10]人工智能技术日新月异，传统治理模式难以跟上时代的步伐，而实验主义治理则具有天然的适配性。实验主义治理可以让科技企业和监管机构等在实验中逐步掌握人工智能的技术性能及潜在风险，通过不断优化治理策略来适应技术的发展变化。而且，实验主义治理可以推动政府部门、科技企业、科研机构、社会公众等多元主体参与其中，整合各方的知识和经验，探索适应人工智能创新发展的治理方案。有鉴于此，本文拟对实验主义治理模式进行全面、系统地研究。

二、实验主义治理何以定位：内涵、机理及渊源

1. 实验主义治理的内涵厘定

实验主义治理是一种以动态实验为核心逻辑，以包容审慎、分类分级、迭代递归为根本理念，以治理的系统性、动态性、适应性为圭臬准绳，适用于人工智能等创新发展领域的新

型治理模式。其核心要义是在多元主体的协同参与下，在安全可控的范围内，按照计划方案进行实验探索，以实验数据和经验成果为依据，动态调整及完善治理策略，不断提升实验对象的安全性、可靠性、可控性，最终实现治理活动科学性和有效性的统一。人工智能社会实验便是人工智能实验主义治理模式的具体体现之一。人工智能社会实验通过长周期、跨学科的实证研究，系统观测人工智能对个人、组织、社会的综合影响，积累实验数据与实践经验，本质上是将实验主义治理的核心理念转化为具体的实践行动，为治理策略的完善提供坚实支撑。[11]社会实验通过具体实践优化实验主义治理的系统设计与框架结构，而实验主义治理则为社会实验提供方法指引和运行逻辑，二者共同推动人工智能治理的科学化发展。

首先，治理主体具有多元性与协同性。实验主义治理强调政府部门、科技企业、科研机构、社会公众等共同参与治理活动。通过多元主体之间的协同合作和优势互补，在实验中可以形成纵向联动、横向协同的治理合力。以粤港澳大湾区生成式人工智能安全发展联合实验室为例，“联合实验室由国家互联网应急中心作为指导单位，广东省委网信办和国家互联网应急中心广东分中心联合牵头，粤港澳大湾区相关部门、企业、高校、科研机构、媒体等共同参与建设。”[12]

其次，治理依据具有灵活性与动态性。实验主义治理要求治理依据具有灵活性和动态性，能够根据人工智能创新发展等实际情况进行不断调整，使得治理依据与现实状况相匹配，以此及时回应数智时代的新问题和新挑战。以阿西洛马人工智能原则为例，该原则涉及保护人类安全、确保人工智能可控性等人工智能发展的基本原则。[13]随着人工智能技术服务的快速发展和应用场景的不断拓展，科技企业根据自身认知和实践经验，对这些原则进行阐释和践行。该原则逐渐形成一套更加具体且完善的治理规则，以此适应人工智能创新发展所带来的各种复杂问题。

再次，治理过程具有积累性和迭代性。实验主义治理是在可控范围内进行实验，不断积累实验数据和经验成果，持续修正治理偏差，推动治理策略从粗放向精准、从模糊向清晰渐进优化。例如，在自动驾驶技术的研发过程中，科技企业通过大量路测来收集数据，相关数据包括多传感器数据、交通场景数据、交通参与者行为数据、驾驶人员操控数据等。随着路测数据的不断积累，研发团队可以根据这些数据对自动驾驶系统在特殊天气下的路况识别、复杂交通场景中的驾驶决策等进行调适，使得自动驾驶系统更加符合实际情况。[14]

最后，治理结果具有科学性和有效性。实验主义治理不仅允许在实验范围内突破现有规则，为人工智能创新发展预留空间，而且通过构建风险防控机制，确保实验活动风险可控。例如，传统诊疗规范是基于医学研究和长期临床经验制定的。医疗人工智能可以通过对大量医疗数据的学习，发现一些传统诊疗规范中未曾提及的治疗方案。实验主义治理允许医疗人工智能在严格监控下在一定范围内超越既有诊疗规范。在实验过程中，对诊疗安全、治疗效果等进行监测，收集数据并进行评估。如果新的治疗方案被证明是有效的且风险可控，那么可以将其逐渐纳入诊疗规范。[15]

2. 实验主义治理的内在机理

实验主义治理的基本逻辑在于实验、评估、优化、再实验的迭代循环。以人工智能监管沙盒制度为例，监管沙盒通过在研发和上市前阶段建立安全可控、动态监管的环境，在人工智能系统提供者与政府部门商定的特定沙盒计划下，在限定的时间内测试人工智能系统的性能、安全性及潜在影响等，之后再将这些人工智能系统投放市场或以其他方式投入使用。[16]在实验阶段，监管沙盒为人工智能系统提供者建构安全可控的测试空间，在限定时间内测试人工智能系统的性能状况及潜在影响。在评估阶段，监管沙盒形成风险管控与效果验证的双轨体系。治理主体可以对重大风险进行监控、识别及应对，可以对测试结果进行全面研判，验证创新方案的可行性与合规性。在优化阶段，治理主体可以基于评测结果调整治理策略，推动

治理规范从原则性框架迈向精细化规则。

实验主义治理的底层逻辑是通过对人工智能系统进行实验探索、科学评估、及时响应，实现治理规则与创新发展的动态适配。人工智能监管沙盒通过受控环境，及时掌控人工智能模型算法、数据信息、技术服务等指标变化，持续关注安全、健康及权利保护领域的潜在风险。治理主体可以通过直接监控、测试报告审核、多方主体反馈等渠道积累全周期数据。若测试中发现无法消解的重大风险，治理主体可暂停或终止测试，即时阻断风险扩散。政府部门可以根据测试结果细化准入退出标准、调整责任豁免条件等，科技企业可以通过年度报告汇总实践成果与经验教训，科研机构可以将成熟经验转化为立法依据或行业标准，从而使得治理规则始终与技术创新同步演进。

实验主义治理的内在逻辑强调通过设定差异化的治理目标，激发多元主体在技术创新与合规实践中的竞争动力。人工智能监管沙盒呈现出目标定制设定、主体多维竞争、成效动态转化等特征。监管沙盒可以为高风险人工智能系统设置全生命周期合规要求，可以为低风险人工智能系统赋予公正透明义务等。在场景细分维度，监管沙盒可以针对不同行业、不同企业，制定个性化基准。通过这种立体化基准体系，监管沙盒为不同人工智能系统提供者提供可量化的竞争标尺，建构跨行业对标、同赛道竞技的良性竞争格局。监管沙盒中的基准竞赛在企业和区域两个层面同步展开，形成多层次创新激励。在企业层面，监管沙盒通过合规认证、市场准入的联动机制，将实验结果转化为竞争优势。在区域层面，通过沙盒清单和经验共享推动区域间竞争。监管沙盒可以通过数据追踪、评估反馈、规则升级的闭环机制，将竞赛成果转化为制度创新。

实验主义治理的关键逻辑是对人工智能创新风险的精准管控与治理效能的持续提升。政府部门、科技企业、科研机构等对监管沙盒内人工智能的创新性、安全性以及社会影响性进行实时追踪。治理内容不仅涵盖模型算法透明度、数据信息安全等技术指标，而且包括对人机关系异化、责任认定模糊、数字鸿沟加剧等非技术维度。监管沙盒为不同的人工智能系统，设置了差异化的预警指标。针对一般预警，监管沙盒启动适应性调整程序，通过指导科技企业优化模型算法参数、完善技术服务方式等修正偏差；对于严重预警，治理主体立即启动终止程序，启动风险处置预案，强制项目退出沙盒。治理主体可以将实验过程中发现的共性问题转化为治理规则的优化建议。

3. 实验主义治理的理论渊源

实验主义治理作为一种强调问题导向、实践检验、动态适应的治理模式，其理论根基在于实用主义哲学。其一，实用主义哲学主张，思想、理论、制度的价值不在于是否符合某种抽象本质或先验原则，而在于其能否有效解决实际问题、指导行动并产生可验证的效果。[17] 实验主义治理强调在局部范围内先试先行，通过收集实验数据、评估实际效果来判断人工智能系统的安全性、可靠性、可控性、公平性，筛选出对具体问题最具解释力和解决力的治理策略。这种效果优先导向在本质上是实用主义在治理领域的延伸。其二，实用主义哲学强调观念到信念、信念到习惯、习惯到行动的创新性功能，创新的结果是从观念外展、创新出来的行动后果。[18] 实验主义治理将治理过程视为一个持续行动与优化迭代的循环。面对人工智能系统创新发展所产生的复杂治理问题，治理主体先基于经验提出初步方案，通过实验将方案落地实施，根据实验结果修正假设、优化方案。这种动态迭代逻辑使得治理过程始终与实践需求保持对话。其三，实用主义哲学主张从疑难情境中发现问题，提出问题解决方案，而情境必须嵌入具体的历史、社会和文化语境中来理解。[19] 实验主义治理强调治理方案的差异性，同一问题在不同阶段、地域、群体，可能需要差异化的策略，因此需通过实验来检验具体方案。

复杂适应系统理论构成了实验主义治理的理论支柱。其一，社会治理系统作为典型的复杂适应系统，由多元主体构成，主体间存在非线性互动，受环境扰动影响显著，导致系统演

化路径难以通过静态模型进行完全预测。[20] 实验主义治理将不确定性作为逻辑起点，通过小范围、多样化的实验项目，在可控范围内测试不同治理方案的实际效果。实验主义治理通过实验探索降低不确定性，实验中暴露的问题、主体的反馈成为破解人工智能系统黑箱的关键，为后续策略调整提供依据。其二，复杂适应系统的涌现性是指，系统整体行为并非个体行为的简单叠加，而是通过主体互动所产生的全新特征。这些涌现特征难以通过顶层设计预先规划，但可能蕴含解决复杂问题的重要线索。[21] 实验主义治理通过集中实验和多元探索为涌现性创造空间。实验主义治理不预设唯一正确方案，而是鼓励不同地区、不同主体基于本地条件开展差异化实验。通过跨实验的观察、比较和交流，这些局部互动可能涌现出超出单个实验预期的整体规律或创新方案。其三，复杂适应系统中的个体或子系统会根据环境变化和其他主体的行为持续调整自身策略。[22] 而实验主义治理将适应性转化为治理优势，并以此构建了迭代递归循环。

反身性现代化理论所揭示的现代化进程中的内在复杂性与不确定性，为实验主义治理提供了理论依据。其一，现代性中风险具有复杂性、危害性、扩散性等，既有治理模式所依赖的静态规则与预设方案难以应对这类动态演化的风险。[23] 实验主义治理通过持续的实验、反馈、迭代来优化治理策略。其二，现代制度需要不断应对自身运作所产生的新问题，现代制度处于持续自我修正状态。[24] 实验主义治理将制度本身视为可调整的框架，通过建立多元主体参与的实验平台，让制度在实践中暴露问题、收集反馈，进而推动制度规则的迭代。其三，反身性现代化中的个体化趋势使得社会需求与风险承受呈现分化特征，统一的治理模式难以适配多样化的诉求。[25] 实验主义治理通过小规模实验、局部性实验等方式，能够在不同场景中测试治理效果，再基于实验经验进行推广或调整。这种灵活性与针对性回应了个体化带来的治理挑战，使治理更贴近个体在反身性社会中的实际处境。

三、实验主义治理何以展开：多元场景与潜在风险

1. 实验主义治理的典型场景

在自动驾驶领域，国家层面布局17个国家级测试示范区，累计开放测试示范道路3.5万多公里，发放测试示范牌照超过1万张，部署智能化路侧单元超过1.1万套。[26] 地方实验成效显著，杭州建成6910平方公里全国最大测试区域，累计发放测试牌照331辆，测试里程近250万公里。[27]2025年12月工信部公布我国首批L3级有条件自动驾驶车型准入许可，开启首批L3级车型附条件准入，极狐、长安纯电动轿车分别在北京、重庆指定路段试点，限定速度与场景边界，积累复杂数据。[28] 自动驾驶领域实验主义治理重点聚焦三个方面：一是法规供给先行，多地出台地方性法规，明确事故责任划分，消解系统接管期间权责难题；[29] 二是安全可控底线，实行附条件准入、安全员值守等机制，通过场景管控、速度可控降低风险，依托第三方机构开展独立评估；三是完善治理规则，地方实践为国家立法积累经验，推动《自动驾驶汽车法》纳入全国人大立法规划。[30]

在医疗健康领域，北京儿童医院与百川智能联合研发的“福棠·百川”儿科大模型，整合了超过300位北京儿童医院知名儿科专家的临床经验和数十年脱敏后的专家高质量病历数据，构建了覆盖儿童常见病、疑难病的立体化知识体系，测试数据显示，该模型诊断准确率达82%，优于主治医师平均水平78%。[31] 深圳罗湖医院集团接入双模态人工智能系统，实现预问诊、辅助诊断、病历撰写全覆盖，医生只需要根据人工智能提供的辅助诊断结合病人的实际情况作出最终的判断，医院下属46家社康中心也全部运用人工智能辅助医疗系统，借此缓解药剂师紧缺等难题。浙江大学医学院附属第一医院开发出的人工智能病理大模型，不仅能够精准识别出疑似病变区域，为医生提供专业参考，而且能够减轻病理科医生工作负担，促进其专业精进。[32] 医疗健康领域实验主义治

理主要围绕两个方向展开：一是多元主体协同，整合医疗机构、科技企业、政府部门、社会公众等多方力量，构建权责适配的协同治理框架，破解医疗领域利益格局复杂性难题。二是全链安全导向，从智能诊断到精准治疗，聚焦预防、诊疗、康养全链条治理，实现医疗质量提升与公共利益保障的统一。

在金融科技领域，自2019年12月北京开启金融科技创新监管试点应用后，仅仅一年多时间，北京、上海、成都、深圳等十个省市和地区已有93项创新试点应用进入监管沙盒。[33]广州金融科技创新试点项目呈现主体多元、技术前沿、场景精准、价值导向鲜明的显著特点，构建事前精准评估、事中动态监测、事后严格核查的全周期风险防控闭环，实践中的机制创新、技术应用、风险防控和服务实体等多方面经验，为后续其他试点城市提供了完整的工作范式与路径参考。[34]深圳市市场监管局、央行深圳市中心支行联合多部门，在预付式经营领域应用数字人民币试点，试点从教培行业延伸至文旅、餐饮、宠物、运动健身等多个行业，在保障资金安全的前提下，降低监管成本，提高交易效能。[35]金融科技实验主义治理形成实验、评估、优化、再实验的治理闭环，实现创新与安全的动态平衡。实验过程强化穿透式监管能力，提升风险管理的前瞻性和精准性，夯实金融科技高质量发展的制度基础。

在公共服务领域，2024年以来，杭州余杭“余智护杭”系统累计流转处置事件37万余件，日均处置500余件；事件平均处置时长缩短30%以上，整体闭环率达98.7%，质检合格率达97%。“余智护杭”系统构建起智能化治理新范式，通过人工智能技术深度融合基层治理场景，实现了从数智治理向智慧治理的跨越升级，有效提升了基层治理的实战效能。[36]浙江台州路桥区“融”易治平台，以跨部门、跨系统、跨层级的方式获取和融合银行账户、法院诉讼、公安接处警等部门数据，设置非法集资行为风险评价、严重性等级划分与预警监测等风控模型，按照违规严重程度建立风险定级规则。“融”易治平台通过融合多源数据实现穿透式监测，及时把风险阻断在萌芽状态，为地方金融风险治理提供可复制经验。[37]实验活动坚守安全可控底线，以实验成果反哺治理策略，彰显实验主义治理在公共服务领域的科学性与有效性。

2. 实验主义治理的核心挑战

首先，实验主义治理依赖多元主体协同发力，但是实践中因治理主体诉求分野、参与惰性、权责失衡等，在协同层面面临一定的困境，制约治理效能。

其一，诉求分野产生协同难题。政府以公共利益为核心，侧重风险防控与规则建构。科技企业追求技术迭代与商业利益，倾向降低实验合规成本。科研机构聚焦学术突破，注重实验创新性与数据完整性。社会公众则关切隐私保护与权益保障。多元主体目标与利益的差异可能引发博弈。例如，政府严格监管与企业创新诉求形成张力，科技企业可能隐瞒实验风险以加快落地，而公众对技术的认知偏差可能引发信任危机。

其二，参与惰性影响实验结果。参与积极性受资源、认知、利益三重因素制约。中小科技企业及基层科研机构因资金、技术资源匮乏，难以参与复杂实验项目。公众受人工智能专业知识壁垒限制，缺乏参与实验设计、监督的能力与渠道。部分社会公众因无直接利益关联，主动参与意愿薄弱。这种参与不均衡导致实验数据偏向优势主体诉求，难以覆盖不同场景与群体需求，削弱治理策略的全面性。

其三，权责失衡引发实验内耗。政府虽拥有监管权但缺乏前沿技术储备，难以精准把控实验细节。科技企业掌握核心技术与实验数据，却可能规避风险责任。科研机构与社会公众话语权较弱，难以影响实验决策。此外，权责分配缺乏动态调整机制，实验过程中出现的新风险难以快速明确责任主体，可能出现相互推诿的情况，既影响实验进度，又埋下安全隐患。

其次，实验主义治理的实验设计与实施是连接理论框架与治理实效的关键环节。然而，该环节面临技术复杂性、场景多样性、目标多元性困境。一是人工智能创新发展使得实验方

案设计陷入科学性与落地性的两难境地。过于追求科学严谨，可能导致实验方案操作繁杂、成本攀升，超出科技企业的承担能力。侧重现实可行的简化方案又可能遗漏人工智能系统可能存在的安全风险，而且单一环节风险失控可能引发连锁反应，进一步提升了风险管控的难度。二是实验代表性不足将制约治理效能的提升。不同领域的技术路径、伦理诉求、风险阈值等差异显著。若实验主体仅聚焦头部企业、单一场景、特定群体，所得结果可能陷入样本偏差，难以覆盖中小企业、边缘场景、弱势群体的需求。而且，人工智能系统迭代速度较快，实验场景时效性较强，某一阶段的实验结果可能因技术更新而失效，难以推广至更多的治理场景。三是实验动态性与灵活性的适配难度较高。实验主义治理的核心是动态优化，但是人工智能实验的数据体量庞大、实时性要求高，如何全面有效分析数据面临技术瓶颈。另外，治理策略的调整需兼顾安全底线与灵活适配，过度追求灵活性可能突破安全可控底线，过于强调规范性又会错失优化时机。

再次，数据与信息是实验主义治理的核心支撑，其质量、流动及解读效能直接决定治理策略的科学性。受到技术特性、利益诉求、制度约束等影响，数据与信息成为实验主义治理实现的关键梗阻。其一，数据信息质量难以精准把控。人工智能实验数据来源广泛且异构显著，涵盖科技企业技术服务数据、政府部门公共服务数据、海量用户行为轨迹数据等，不同来源数据在采集方式、数据特征、价值密度、敏感程度等存在较大差异，易出现数据缺失、冗余重复、逻辑偏差、格式混乱等问题，给数据处理带来较大挑战。部分主体受商业利益、科研产出等诉求驱动，可能存在篡改数据、选择性筛选数据等行为，为抢占技术落地先机或追求实验结论理想，通过技术手段修饰数据指标，导致数据真实性受损。其二，数据信息流动存在双重矛盾。一方面，多元主体存在数据信息壁垒，如企业因商业机密考虑而不愿公开核心实验数据，政府与科研机构的信息披露也受诸多限制，导致实验数据信息碎片化，难以形成治理合力。另一方面，数据信息共享与隐私保护、数据安全等存在冲突。过度强调共享可能会泄露企业核心技术及用户隐私，严苛的安全管控又会阻碍信息流动，现有制度难以平衡二者关系。其三，数据信息解读存在短板。实验数据海量且复杂，需要融合多学科知识解读，而当前专业分析团队稀缺，多数机构缺乏跨领域复合型人才。此外，中小企业及基层科研机构因资金、技术投入有限，数据分析能力相对薄弱，导致实验数据价值无法充分释放。

最后，权利保障是实验主义治理的合规基础，可是实验活动在权利主体界定、权利冲突协调、权利救济途径等面临多重挑战。在权利主体界定方面，实验活动涉及技术研发者、服务提供者、实验参与者、社会公众等多元法律主体，部分主体权利范围难以依据现行法律进行精准界定。在权利冲突协调方面，不同主体权利诉求的博弈，缺乏明确法律规则指引。科技企业可能以保护商业秘密为由隐瞒实验关键信息，与公众基于知情权、监督权的合理诉求形成冲突。实验活动往往简化风险告知、压缩知情同意时间。这导致实验效率诉求与参与者权利保障可能存在龃龉。在权利救济途径方面，实验的技术复杂性使权利损害具有隐蔽性、滞后性，如算法偏见导致的歧视性损害、算法黑箱导致的举证难问题。个体面对科技企业时，存在取证周期长、举证成本高等问题，现有救济体系难以充分保障权利主体合法权益。

四、实验主义治理何以实现：目标引领与制度保障

1. 优化实验主义治理目标

实验主义治理的目标是创新可续、风险可控、价值可塑的三位一体目标体系。该体系不仅立足于人工智能创新发展的时代趋势，而且锚定人类社会以人为本、智能向善的核心诉求。实验主义治理通过多元协同治理体制与动态灵活治理机制，推动治理目标的全面实现。

创新可续目标聚焦于释放技术活力与维持治理弹性的平衡。技术活力的本质是模型创新、

算力突破、数据赋能的协同共进，其贯穿人工智能从技术原理到产业应用的全生命周期。[38]三者的协同发力，共同构成人工智能技术活力的完整体系，推动弱人工智能向强人工智能的演进。而治理弹性则是应对人工智能技术不确定性的关键。治理弹性依托实时监测、动态评估以及快速响应等，能够及时发现人工智能创新发展中的安全风险，通过调整治理策略，精准消解各类隐患。治理弹性不仅可以避免刚性治理对创新的抑制，而且可以防范技术无序发展引发的安全问题。

风险可控目标的核心在于通过实验探索构建安全、可靠、可控的闭环治理体系。[39]人工智能安全风险具有难以感知性、高度关联性、迅速衍变性等。实验主义治理摒弃既有治理的静态僵化逻辑，通过实验、评估、优化、再实验，精准识别各种风险与防治盲区。实验主义治理既为治理策略提供了实践检验的场景，又能根据技术发展动态调整防控重点。实验主义治理贯穿风险识别、评估、干预、反馈的全链条。在风险识别阶段，依托技术监测与多元主体协同，全面捕捉技术服务中的潜在隐患；评估阶段建立科学的风险分级标准，明确不同场景下的防控举措；干预阶段采取精准化管控措施，在不抑制创新活力的同时，及时阻断风险扩散；反馈阶段则将实践中的治理经验转化为制度优化的依据，形成治理策略与创新发展的动态适配。

价值可塑目标强调，通过实验主义治理将人类福祉、公平正义、智能向善等核心价值深度嵌入人工智能技术研发与应用推广全过程，进而实现技术工具理性与价值理性的辩证统一。[40]技术研发过程可以建立伦理审查机制，对可能危害人类权益的研发方向进行前置管控，让技术从源头就扎根于人类福祉的土壤。应用推广过程可以建立应用场景的动态监测机制，及时发现技术应用中可能出现的隐私泄露、权益侵害等问题。此外，不同国家基于文化传统与社会制度形成差异化的价值偏好，通过共享人工智能治理经验，可以在尊重多元价值的基础上寻求共识性准则。

2. 完善实验主义治理体制

第一，在实验主义治理多元主体中，政府发挥着关键作用。[41]作为统筹者，政府牵头构筑实验主义治理框架。一方面，政府可以制定人工智能领域的宏观战略与发展规划，明确阶段性目标与重点实验方向，为治理主体提供清晰的指引。另一方面，政府可以健全实验主义治理机制，整合科技、工信、广电等多领域资源，对实验全流程进行统筹管理。作为引导者，政府推动多元主体深度协作。政府可以引导科技企业积极参与实验，通过项目补贴、税收减免等政策工具，激励科技企业将研发资源向实验项目倾斜，推动其在创新发展中实现商业利益与社会福祉的平衡；政府可以助推科研机构与科技企业建立协同创新机制，支持科研人员深度参与企业实验，加速科研成果向产业转化；政府还可以引导社会公众充分发挥监督与反馈职能，畅通公众参与实验监督的渠道，系统收集并吸纳公众意见建议，确保实验更契合社会公共需求。作为保障者，政府筑牢实验活动的支撑体系。在制度层面，政府为实验开展提供坚实的法律依据，制定并完善实验计划，明确权利义务边界、数据使用准则、风险责任划分等，维护良好的实验秩序。在资源层面，政府加大资金投入力度，降低实验参与门槛与运行成本，为多元主体广泛参与实验创造有利条件。

第二，科技企业在多元协同治理体系中具有核心作用。[42]作为实验项目的承担者，科技企业凭借技术研发优势与场景落地能力，主导人工智能系统的实验设计与执行。例如，科技企业可以在技术研发中嵌入数据安全、知识产权保护等合规要求；可以对人工智能系统风险进行跟踪评估，建立风险应急预案；可以组建科技伦理审查委员会，对人工智能系统的风险等级、潜在影响等进行前置评估。作为实验活动的枢纽者，科技企业可以主动向社会公开人工智能系统的基本情况、安全风险及治理措施，消除公众对人工智能创新发展的认知壁垒；可以吸纳科研机构参与人工智能实验项目，共建研究平台与人才培养基地；可以建立与监管部门的常态化沟通机制，主动报送企业合规情

况，配合监管检查与执法行动。作为实验反馈的汇聚者，科技企业可以向政府部门提供技术咨询与行业报告，助力政策的科学性与前瞻性；可以依托技术优势与实验经验，参与国家标准、行业标准等起草工作；可以设立公众投诉渠道与意见征集平台，接受社会公众对产品与服务的监督。

第三，科研机构是理论创新的策源地、协同治理的推动者以及实验成效的观察者，其作用贯穿实验主义治理的全流程。[43]科研机构可以为治理实验提供系统性知识框架。通过对人工智能技术原理、安全风险、社会影响等深度研究，科研机构为政府制定实验政策法规、企业设计实验方案提供理论依据，确保治理探索的科学性与合规性。科研机构作为协同治理的推动者，可以搭建多元主体的对话平台。例如，组织科技企业、专家学者、社会代表共同参与实验设计，破解单一主体认知局限导致的治理偏差。更为重要的是，科研机构以独立第三方身份，客观评估政府监管策略、企业实验方案等实际效果。科研机构可以将实验成果转化为政策法规修订建议，通过国际学术合作分享本土实验成果，为全球人工智能治理相关规则的形成提供科研支持。

第四，在多元协同治理体系中，社会公众是实验方向的锚定者、实验活动的参与者以及实验成效的评判者。社会公众对信息茧房、隐私泄露的担忧可以转化为具体治理目标，推动实验方案纳入可解释性、知情同意等核心指标，确保实验探索始终围绕人类福祉展开，避免技术发展与社会需求脱节。此外，社会公众亲身参与治理实验的场景测试，提供不可替代的实践反馈。社会公众通过社会舆论、消费选择等方式对治理实验进行外部约束。其对算法歧视、技术滥用等问题的关注，可以形成倒逼机制，促使科技企业规范实验流程、政府强化监管职责。社会公众通过对不同实验方案的反馈差异，有助于筛选出最契合社会共识的治理模式，并为治理规则的形成提供基于公众认知的参考样本。

3. 健全实验主义治理机制

首先，健全风险管理机制是人工智能实验主义治理的重中之重。人工智能实验前期，应当结合不同行业领域的现实状况、安全需求、法律法规等，明确人工智能实验的边界与禁区，制定差异化的实验方案；组织专家学者等对人工智能实验开展多维度风险论证，依据影响范围、危害程度等将实验风险划分为不同等级，并且针对各级风险制定具体的防控预案；强化合规性审查，完成必要的实验备案审批流程。人工智能实验过程实行动态管控与宣传引导。对高风险应用先开展小范围实验，实时监测应用效果与风险状况，收集实验反馈，完善技术方案和管理措施，待风险可控后再逐步扩大实验范围，避免大规模实验导致风险失控。搭建实时监测平台，对实验运行状态、参与者反馈等进行全方位追踪。一旦发生重大危险，立即启动防控措施。加强实验参与者与社会公众的引导，通过通俗易懂的形式普及人工智能系统的技术原理、应用价值及风险边界，保障实验参与者的知情权与选择权，提升社会公众对人工智能技术的接受度与信任度。

其次，构建适配实验主义治理的信息沟通机制，实现信息在治理主体之间的有序流动。在正式渠道方面，可以建立实验备案与公示制度，科技企业在实验启动前，提交详细方案并通过官方平台予以公示，接受社会监督；可以通过季度报告、年度总结等形式，系统披露实验进展、数据指标和治理成效；可以设立高风险实验专项沟通通道，建立政府部门与科技企业的对接机制，确保风险信息第一时间传递。在非正式渠道方面，可以搭建行业交流平台，通过专题研讨会等形式促进科技企业与科研机构之间的经验共享；可以依托官方网站、社交媒体等途径，及时收集公众诉求。此外，专业信息披露需要兼顾专业性与通俗性，让非专业主体能够理解实验核心内容。风险信息实行分级披露制度，对一般风险进行常规公示，对重大安全风险进行专项通报。针对实验过程中出现的新问题新风险，及时组织相关主体开展专题会商，调整沟通重点和信息披露内容。建立信息审核机制，由专业团队对披露信息的真实

性、完整性以及合规性进行审核，杜绝虚假信息、误导性陈述和遗漏上报问题等。

再次，权利保障机制的重点在于明确权利边界、畅通权利表达、完善权利救济，确保技术创新不偏离人类福祉。对于权利边界而言，科技企业必须以清晰、易懂的方式告知参与者与实验有关事项，由参与者在充分知情的前提下自愿、明确作出同意。针对高风险实验项目，应当取得参与者单独同意或书面同意。参与者在实验的任何阶段，均有权无理由退出实验。参与者有权知悉实验的实际进展情况，核查实验是否严格按照实验方案执行。若实验方以有偿方式招募参与者，双方应当通过合同明确报酬标准、支付方式与支付期限等。对于权利表达而言，在实验启动阶段，通过组织听证会等形式，收集实验相关方权利诉求，并将其转化为实验方案。在实验实施阶段，设置实时反馈端口，允许权利主体对实验活动提出异议。在实验评估阶段，引入第三方机构对权利实现程度进行测评，将测评结果作为下一轮实验调整的依据。对于权利救济而言，由政府部门受理人工智能领域的权利侵害投诉，简化流程快速处置；完善司法诉讼机制，配备专业审判人员，明确举证责任分配规则，缓解参与者因信息不对称导致的举证难题；引入调解仲裁机制，高效化解技术门槛高、争议标的小的权利纠纷。

最后，责任豁免机制是激励科技企业参与实验的关键制度设计，其通过明确实验过程中的责任范畴，在保障社会公共利益的前提下，为实验探索预留空间。[44] 责任豁免的适用范围应当基于实验类型精准划定。针对技术验证类实验，可以对符合预设标准的科技企业适用合规豁免，即只要实验流程严格遵循事前报备的方案且未突破风险控制的底线，对实验中产生的非预期技术偏差免于追责。对于应用推广类实验，豁免范围需要限定于技术固有局限导致的轻微损害，对于科技企业因故意或重大过失引发的严重后果仍然需要承担责任。此外，科技企业只有满足三重前提，才能申请责任豁免：一是完成风险备案，在实验启动前提交详尽的风险评估报告，明确潜在损害的类型与范围；二是执行防控方案，在实验过程中严格落实预设的风险防控措施；三是保持过程透明，定期向政府部门报送实验进展数据，接受第三方机构的监督。相关条件应当根据技术迭代情况进行相应调整，避免僵化的条件设定抑制实验创新。

结　　语

人工智能创新发展正在实现从内容生成向智能行动的关键跨越。智能体作为技术变革的核心，重塑了人机协作的底层逻辑。多模态融合技术的持续迭代与推理能力的突破提升，推动人工智能逐步形成类人感知与思考的综合能力。人工智能与经济社会各行业各领域广泛深度融合，形成跨界融合、共创共享的智能经济和智能社会新形态。然而，人工智能既有治理模式却面临着技术特性与治理逻辑的结构性张力，法律治理的滞后性、设计治理的僵化性、风险治理的盲目性，均难以适配人工智能技术快速迭代、场景多元分化、风险动态演化的发展趋势。如何应对人工智能技术带来的各种风险和复杂挑战，促进人工智能技术造福于人类，成为亟需回答的重要问题。实验主义治理作为一种动态平衡、协同演进的人工智能治理新模式，契合人工智能的技术特性与发展需求。为了构建人工智能实验主义治理体系，我国需要确立创新可续、风险可控、价值可塑的三位一体治理目标，设立政府统筹、企业主导、科研支撑、公众参与的多元协同治理体制，健全以风险管理、信息沟通、权利保障、责任豁免为重点的治理机制，在推进国家治理体系和治理能力现代化的同时，确保人工智能技术朝着有利于人类文明进步的方向健康有序发展。

［参 考 文 献］

[1] 闫瑞峰. 算法设计伦理治理的立场、争论与对策 [J]. 自然辩证法通讯, 2023, 45（6）: 1–9.

[2] 杨雪冬. 全球化、风险社会与复合治理 [J]. 马克思主义与现实, 2004,（4）: 61–77.

[3] 丁晓东. 人工智能风险的法律规制——以欧盟《人工智能法》为例 [J]. 法律科学（西北政法大学学报）, 2024,

42（5）: 3–18.
[4] Sabel, C., Zeitlin, J. *Experimentalist Governance*[M]. Oxford: Oxford University Press, 2012, 169–184.
[5] Huitema, D., Jordan, A., Munaretto, S., et al. 'Policy Experimentation: Core Concepts, Political Dynamics, Governance and Impacts'[J]. *Policy Sciences*, 2018, (51): 143–159.
[6] 'Regulatory Sandbox'[EB/OL]. https://www.fca.org.uk/firms/innovation/regulatory-sandbox. 2025–04–05.
[7] Burd, J. 'Regulatory Sandboxes for Safety Assurance of Autonomous Vehicles'[J]. *University of Pennsylvania Journal of Law and Public Affairs*, 2021, (12): 194–226.
[8] 'National Bioengineered Food Disclosure Standard'[EB/OL]. https://www.federalregister.gov/documents/2018/12/21/2018-27283/national-bioengineered-food-disclosure-standard. 2025–04–05.
[9] 坚持自立自强 突出应用导向 推动人工智能健康有序发展[N]. 人民日报, 2025–04–27（1）.
[10] 国务院关于深入实施“人工智能+”行动的意见[EB/OL], https://www.gov.cn/zhengce/content/202508/content_7037861.htm. 2025–09–05.
[11] 国家新一代人工智能创新发展试验区建设工作指引（修订版）[EB/OL], https://www.most.gov.cn/xxgk/xinxifenlei/fdzdgknr/fgzc/gfxwj/gfxwj2020/202012/t20201224_171987.html. 2026–01–29.
[12] 叶青. 粤港澳大湾区生成式人工智能安全发展联合实验室成立[N]. 科技日报, 2025–09–16（2）.
[13] 商瀑. 从“智人”到“恶人”: 机器风险与应对策略——来自阿西洛马人工智能原则的启示[J]. 电子政务, 2020, 17（12）: 69–76.
[14] 陈力、黎晓奇、王赫然. 自动驾驶汽车技术的应用风险与法律规制研究[J]. 中国科技论坛, 2025, 41（4）: 22–30.
[15] 汪琛. 医疗人工智能伦理治理的问题、困境与求解[J]. 科学学研究, 2025, 43（2）: 414–422.
[16] 徐磊. 欧盟人工智能监管沙盒的制度构造与经验启示[J]. 科技管理研究, 2025, 45（9）: 21–30.
[17] 宋斌、闫星宇. 杜威实用主义哲学的现实审视[J]. 求索, 2013,（6）: 92–95.
[18] 卢德平. 皮尔士实用主义准则的意义问题[J]. 哲学分析, 2023, 14（4）: 38–47.
[19] 陈哲、马雷. 情境、探究与问题——杜威实用主义问题方法论初探[J]. 科学技术哲学研究, 2023, 40（5）: 64–68.
[20] 张立荣、方堃. 基于复杂适应系统理论（CAS）的政府治理公共危机模式革新探索[J]. 软科学, 2009, 23（1）: 6–11.
[21] 毛征兵、陈略、范如国. 中国开放经济系统及其发展模式的机理研究——基于复杂适应系统范式的解析[J]. 经济与管理研究, 2021, 42（1）: 16–39.
[22] 刘洪. 组织结构变革的复杂适应系统观[J]. 南开管理评论, 2004, 7（3）: 51–56.
[23] Farrugia, D. 'Addressing the Problem of Reflexivity in Theories of Reflexive Modernization: Subjectivity and Structural Contradiction'[J]. *Journal of Sociology*, 2015, 51(4): 872–886.
[24] Williams, M. 'Security Studies, Reflexive Modernization and the Risk Society'[J]. *Cooperation and Conflict*, 2008, 43(1): 57–79.
[25] Argyrou, V. 'Reflexive Modernization and other Mythical Realities'[J]. *Anthropological Theory*, 2003, 3(1): 27–41.
[26] 何宇. 场景扩容 生态创新——产业链上下游共话智能网联汽车发展[N]. 中国交通报, 2025–11–06（8）.
[27] 市投资促进局. 6910平方公里！杭州建成全国最大智能网联车辆测试应用“田”[EB/OL], https://tzcj.hangzhou.gov.cn/art/2025/3/20/art_1621408_58894502.html. 2026–01–18.
[28] 装备工业一司. 我部许可两款L3级自动驾驶车型产品[EB/OL], https://www.miit.gov.cn/jgsj/zbys/qcgy/art/2025/art_6e3e04a540a143ba9ce616c40904033b.html. 2026–01–18.
[29] 汪全胜、宋琳璘. 无人驾驶汽车与我国道路交通安全法律制度的完善[J]. 中国人民公安大学学报（社会科学版）, 2020, 36（3）: 107–115.
[30] 唐孝忠. 应尽快制定自动驾驶汽车法[N]. 重庆法治报, 2025–03–07（3）.
[31] 北京经济技术开发区管理委员会. 全球首个儿科大模型落地荣华医院[EB/OL], https://www.beijing.gov.cn/ywdt/gzdt/202506/t20250617_4115339.html. 2026–01–18.
[32] 辅助诊断、AI问答、健康管理……AI正重塑医疗服务链条[EB/OL], https://www.cac.gov.cn/2025-03/17/c_1743916011485636.htm. 2026–01–18.
[33] 李冰. 十地公示93个金融科技“监管沙盒”试点项目银行参与度超九成[N]. 证券日报, 2021–06–15（B1）.
[34] 广州资本市场金融科技创新试点首批13个项目退出评审工作顺利完成[EB/OL], https://finance.sina.com.cn/roll/2026-01-13/doc-inhheixp9062881.shtml. 2026–01–18.
[35] 深圳率先应用数字人民币破解预付式经营监管难题预

付式消费再也不怕商家“跑路”[N]. 深圳晚报, 2023–06–27（A4）.

[36] 中共杭州市余杭区委社会工作部. 以AI之笔绘就基层智治新画卷 [EB/OL], https://www.rmlt.com.cn/2025/1121/743091.shtml. 2026–01–18.

[37] 建立金融风险全链条治理新体系 [EB/OL], http://www.luqiao.gov.cn/art/2022/8/12/art_1229663543_58938733.html. 2026–01–18.

[38] 米加宁. 生成式治理: 大模型时代的治理新范式 [J]. 中国社会科学, 2024,（10）: 119–139.

[39] 徐磊. 发展与安全并重: 生成式人工智能风险的包容审慎监管 [J]. 理论与改革, 2024, 40（4）: 67–83.

[40] 冯子轩. 生成式人工智能应用的伦理立场与治理之道: 以ChatGPT为例 [J]. 华东政法大学学报, 2024, 37（1）: 61–71.

[41] 王小芳、王磊. “技术利维坦”: 人工智能嵌入社会治理的潜在风险与政府应对 [J]. 电子政务, 2019, 16（5）: 86–93.

[42] 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.

[43] 葛春雷、裴瑞敏、张秋菊. 德国科研机构协同创新组织模式研究 [J]. 中国科学院院刊, 2024, 39（2）: 345–357.

[44] 杨丰一. 人工智能监管沙盒法律责任豁免制度研究 [J]. 中国特色社会主义研究, 2025, 38（1）: 98–112.

[责任编辑　李斌]