文章编号:1003 - 2053(2026)05 - 0910 - 11

静态到敏捷:人工智能监管沙盒治理机制研究

吕 悦[1,3],陈 旭[1],彭子璇[2],许 愉[1]

(1. 上海交通大学国际与公共事务学院,上海 200030; 2. 东南大学公共卫生学院,江苏南京 210009;
3. 上海工程技术大学,上海 201600)

摘 要:传统监管模式已难以有效应对人工智能技术所带来的变化。源于金融科技领域的监管沙盒扩展至人工智能领域,成为平衡安全与发展的工具,已被全球多个国家所采纳。当前,人工智能监管沙盒的机制设计尚处于探索阶段,在法律授权、规制边界及治理机制等关键环节上,各国实践表现出路径分化。本文通过梳理人工智能监管沙盒的起源、演进路径及国际应用实践,以"实验主义治理"作为指导人工智能监管沙盒构建与运行的理论框架,比较全球 9 个国家人工智能监管沙盒运行流程,提炼出框架目标设定、实验自主执行、监督与反馈、评议与调整 4 项机制,解释监管沙盒在推动动态调整、多方协作与政策创新方面的适用性。监管沙盒通过对政策目标的持续优化,推动监管主体、创新主体与消费主体在实际运行中通过多方协作和实时动态的反馈渠道优化监管框架,实现技术创新与合规之间的良性互动。发展人工智能监管沙盒可从推进分领域监管沙盒体系化布局、建立可量化的申请标准与评分机制、构建协同共治的合作机制架构、探索跨国互认与标准协调机制着手,提升中国人工智能治理能力,促进国内外监管体系的协同与创新,应对全球技术带来的治理挑战。

关键词:技术治理;监管沙盒;治理改革;实验主义治理;比较研究

中图分类号:F49 **文献标识码**:A

随着人工智能技术的发展,各国在推动人工智能场景应用的同时,也引发了诸如数据隐私、算法黑箱等治理难题[1]。基于固定规则的静态监管模式已难以应对人工智能技术的复杂性和快速变化,易引发监管滞后和治理失效等问题[2]。监管沙盒作为一种创新型监管工具,被广泛应用于新兴技术"试点"测试。通过允许企业在受控环境中测试新技术,并接受监管机构的实时监督与反馈,监管沙盒实现了"实验主义"方式,对科技产品实施动态灵活的治理[3],已扩展至数据要素治理[4]、个人信息保护[5]和人工智能治理[6]等多个新兴技术领域。

欧盟《人工智能法案(AI Act)》要求成员国设立人工智能监管沙盒,并在不同治理重点和业务领域开展试点探索。人工智能监管沙盒正迅速成为各国平衡安全与创新的重要制度工具。尽管沙盒工具的实践在持续推进,其理论研究仍处于滞后状态。目前已有研究集中在法理层面[6],关于人工智能监管沙盒在治理机制、运行逻辑以及政策演化路径方面的系统研究仍较薄弱。AI Act 作为一项宏观法律框架,仅提供原则性要求,具体实施需由各国根据本国实际情况进行沙盒机制设计,这使得各国在监管沙盒的运行路径、激励安排和数据治理模式等方面呈现出明显差异。在此背景下,亟需引入能够解释和指导动态制度演化的理论工具,并从已有实践中提炼出监管沙盒的一般运行机制,为实践提供可操作的制度指引,以应对人工智能技术所带来的高度不确定性和跨域治理挑战。

本文引入实验主义治理框架,通过比较分析视角,系统梳理各国监管重点与治理模式的异同,力图揭示人工智能监管沙盒背后的机制设计原理。文章的主要贡献体现在三方面:(1)在理论层面,构建"框架目标设定—实验自主执行—监督与反馈—评议与调整"四阶段的实验主义治理分析框架,为理解监管沙盒制度演化提供结构性分析与理论解释;

收稿日期:2025 - 04 - 24;**修回日期**:2025 - 08 - 14

基金项目:浙江省软科学研究计划项目(2024C35099)

作者简介:吕悦(1990 -),女,博士研究生。陈旭(1996 -),男,博士研究生,通讯作者,E - mail:cx19961225@ sjtu. edu. cn。彭子璇(1995 -),女,讲师。许愉(1996 -),女,博士研究生。

(2)在实践层面,通过比较九个具有代表性国家的人工智能监管沙盒模式,归纳出三种主流治理取向与执行路径,揭示其在制度类型、测试对象、评估机制等方面的异同,并进一步抽象出人工智能监管沙盒的一般运行流程;(3)在政策建议层面,提出适合中国情境的本土化路径与跨国协同策略,回应当前在沙盒制度建设中面临的标准模糊、责任不清与激励不足等关键问题。本文不仅拓展了人工智能治理的理论研究,也为构建具有前瞻性、灵活性与协同性的中国人工智能监管沙盒体系提供理论支撑与制度借鉴。

1　监管沙盒:起源、发展与实践

1.1　监管沙盒的起源与发展

沙盒本是计算机术语,指搭建安全机制,并为运行程序提供隔离环境[7]。沙盒机制常用于测试一些风险难以判定的程序,通过提供一个受监视和控制的环境,在确保环境真实、方法准确的同时,不对"盒外"环境造成影响,从而保障安全实验。

英国金融行为监管局(FCA)率先提出监管沙盒,为金融科技公司在真实市场中测试创新产品提供受控环境[8]。监管沙盒的主要作用是测试在现有监管结构之外新的经济、制度和技术方法[9],通过监管机构和市场参与者之间的交流互动,使得监管框架变得更加有弹性,从而实现创新与风险的平衡[10]。

监管沙盒在多个层面带来了积极影响。首先,监管沙盒通过降低监管不确定性,使企业更容易获得融资。其次,沙盒机制有助于降低市场准入障碍,减少行政和交易成本,英国 FCA 沙盒计划支持的公司,其产品上市速度比常规流程快约 40%[11]。最后,监管沙盒提供了监管政策制定和监管流程的反馈,增强了监管机构与行业之间的信息互通与知识共享。

1.2　监管沙盒的特征

通过梳理各国发布的监管沙盒的操作指南以及"经济合作与发展组织(OECD)人工智能政策观察站"的相关政策文件,总结出监管沙盒具备三项核心特征。首先,它具有临时性设置特征,监管沙盒为新兴技术提供一个灵活的监管环境,使创新产品或服务能够在受控的条件下进行测试和评估。其次,监管沙盒采用实验方法,通过反复测试和调整,优化监管框架与技术实践的契合度。最后,监管沙盒强调多方协作,鼓励监管机构、技术开发者及其他利益相关方共同参与,增强治理模式的科学性和有效性。

人工智能监管沙盒机制中,主要涉及监管主体、创新主体与消费主体三类利益相关方。监管主体通过提供政策引导与动态调整监管策略,实现对新兴技术的精准治理;创新主体在受控环境中开展技术或商业模式的实验,通过灵活测试与监管互动,降低合规不确定性带来的进入壁垒;消费者作为早期参与者,在促进产品迭代与优化的同时,其反馈也增强了技术应用的可信度与合规保障。三方在沙盒机制中形成良性互动与协同共治,推动技术创新与监管体系的有机融合。

表 1　监管沙盒对不同利益相关者的积极影响

Table 1　Positive impacts of regulatory sandboxes on different stakeholders

监管主体	创新主体	消费主体
提供数据支持与实践依据 推动监管灵活适应新兴技术 促进监管机构与市场合作 动态评估并优化监管框架	简化审批,缩短市场推出时间 降低监管不确定性,避免技术被禁止 收集监管反馈,优化产品制定 提升企业融资机会 降低市场准入壁垒,促进公平竞争	促进安全的新产品引入 增强产品和服务的可获取性

资料来源:OECD (2023), "Regulatory sandboxes in artificial intelligence", *OECD Digital Economy Papers*, No. 356, OECD Publishing, Paris, https://doi.org/10.1787/8f80a0e6-en。

1.3　监管沙盒的种类

监管沙盒主要存在授权式沙盒、沙盒保护伞和虚拟沙盒三种类型。授权式沙盒面向已获得监管许可的企业,允许其在受控环境中测试产品,合规性强,有助于增强市场信任,但准入门槛和运营成本较高;沙盒保护伞为尚未取得直接授权的企业提供

“借壳”测试路径,借助非营利性机构的资质开展实验,降低创新门槛;虚拟沙盒通过模拟环境和共享数据支持测试,尤其适用于资源有限的小微企业,虽然便利性较高,但测试结果与现实场景存在差距。

多样化的人工智能监管沙盒类型为中小企业和初创公司提供了灵活的测试环境,帮助其在较低合规成本下进行创新和产品测试。企业可根据自身需求选择适合的沙盒类型,推动创新成果的落地,降低市场准入门槛。人工智能监管沙盒处于不断演进之中,但从 AI Act 以及英国、法国、德国等国的实践来看,中小企业和初创公司已成为监管支持的重点对象。

2 实验主义治理:分析框架与实践路径

2.1 人工智能监管的范式回顾

人工智能治理存在基本共识也面临诸多争议。一方面,传统的治理模式难以应对人工智能技术的复杂性已形成基本共识[12]。传统治理准则固有的线性特征与渐进式发展模式,导致我国新一代人工智能治理体系面临时效性滞后问题,引发治理效能弱化和功能失调等问题[13]。政策更新的速度难以跟上技术迭代的节奏,现有监管框架无法及时适应新兴应用场景[14]。另一方面,尽管现有治理模式的不足被普遍认可,但尚未探索出适合人工智能技术发展的治理模式[15,16]。

敏捷治理作为新型监管范式,其动态响应特性与人工智能的发展需求契合[17]。武振国[18]认为欧盟的实验主义治理路径通过多主体的试错学习,能够快速应对人工智能技术的迅速变化和不确定性。苏峻[19]提出在具体应用场景开展人工智能社会实验,有助于采集实验数据并优化治理框架,渐进优化治理结构。关于人工智能风险的治理模式,存在规制型治理、创新型治理、自治型治理等治理模式的争论,但敏捷治理和实验主义治理已经成为当前人工智能治理中的主流观点[16]。

关于这两方面的研究仍存在两方面的不足。一方面,从治理模式的角度来看,敏捷治理强调快速响应,但往往偏向于事后治理;而实验主义治理属于事前治理,但其长周期的实验过程常常难以避免对外部环境的干扰,可能人为减缓先进技术的应用。另一方面,现有的治理路径和政策工具箱尚不明确,治理模式仍停留在理论构想层面,缺乏系统性和操作性,导致在实际应用中难以构建有效的治理框架和实施路径。尽管已有学者提出基于治理对象的分类治理思路和治理工具箱匹配构建[15,20],并将监管沙盒引入治理实践[21],但多数研究停留在法律条文内容的分析,这种做法的实践指导意义有限。本文通过分析监管沙盒政策文件和相关材料,提出监管沙盒是一种实验主义治理工具,在实践中弥合了实验主义治理与敏捷治理差异。

2.2 实验主义治理:理论溯源与中国本土化演进

实验主义治理是一种动态、协作和多层次的政策制定方法,通过“学习差异”的过程,促使不同参与者根据实际经验来调整和优化政策[22]。实验主义治理体系中,反馈和讨论推动政策调整,取代了传统的上级引导,从而为政策部门的协作与发展带来新的理解。许多学者呼吁当前的科技治理应转向实验主义治理模式[23]。

中国的政策实践一直蕴含着实验主义治理精神。改革开放以来,中国的政策实践逐渐形成了“摸着石头过河”的渐进式试点机制,成为中国特色实验主义治理的典型代表[24]。试点机制突出了政策实施中的灵活性与适应性,通过在小范围内先行试点,积累经验并逐步推广,通过不断的实验、反馈、迭代与调整,政策逐渐实现优化[25]。监管沙盒在受控环境中测试新技术和商业模式,被认为是实验主义治理的最新演进[18]。

2.3 实验主义治理框架

实验主义治理被视为一种多层次、循环式的治理架构,核心在于通过反馈机制与动态调整,持续优化政策目标,涵盖目标设定、实验自主执行、监督与反馈、评议与调整四个关键环节[23],使其在不确定环境中不断修正政策方向[24]。人工智能监管沙盒的运行机制与上述四个环节在制度结构与实践逻辑上高度契合。监管沙盒通常以明确的政策目标为导向,设定测试边界与基本规范;企业在受控环境中自主开展技术实验,监管机构通过实时监督与数据收集实施动态反馈,最终基于实验结果进行评议与调整,推动机制的不断迭代与优化。这一运行过程展现了实验主义治理所强调的“实验—反馈—迭代—调整”的动态治理逻辑。

(1)框架目标设定

在实验主义治理框架中,设定框架目标是第一步,为政策实施提供方向推动。框架目标通常较为宽泛,强调灵活性与适应性,以便为各级主体提供指

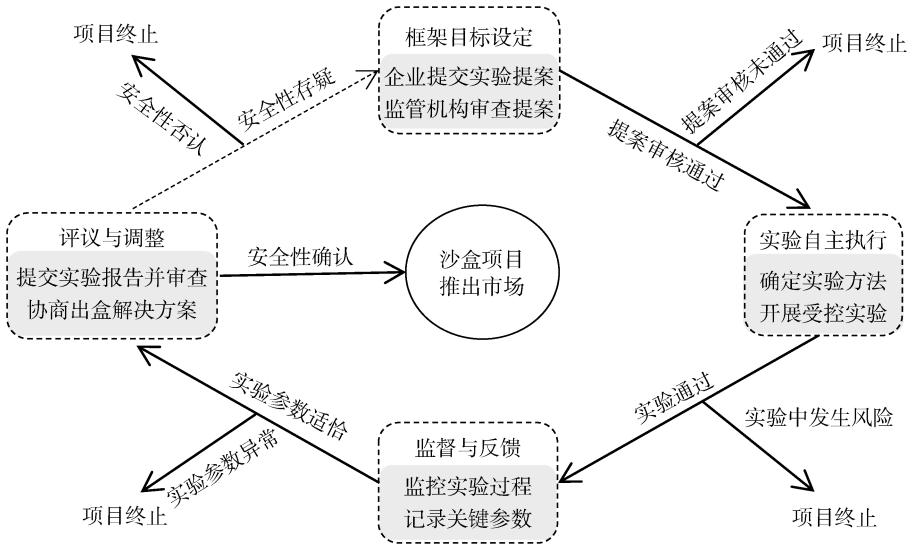

图 1　实验主义治理流程与监管沙盒阶段对照图

Figure 1　Experimentalist governance process versus regulatory sandbox stages

导,保持调整空间,以适应快速变化的环境和不断发展的技术[22]。框架目标设定的过程常要求对监管目标进行宏观把握,识别潜在风险并制定预防措施。

在英国信息专员办公室(ICO)领导的隐私监管沙盒项目中,ICO 将用户隐私和数据安全等核心原则作为框架目标,在此基础上开展前置风险评估,探讨算法可能引发的偏见及数据滥用等问题。通过将关键风险纳入初期讨论并制定相应的合规策略,沙盒在后续运行中得以迅速聚焦指引重点并实现动态调整,确保了监管目标与技术创新的协调推进。

(2)实验自主执行

实验自主执行是框架目标设定后的关键环节。治理主体会将宽泛的目标与原则下放,以便根据具体情境灵活落实政策。通过“自下而上”的反馈与小规模测试,基层单位可获取真实的环境数据及风险信息,进行方案的不断打磨与调整[27]。地方或行业参与者在执行过程中积累的实践经验会回传至上级或平台方,为后续的评估与修订提供依据。如此形成的动态循环,体现了实验主义治理“在实验中学习、在反馈中迭代”的核心理念。

在监管沙盒中,该环节对应项目实验阶段。监管沙盒创新主体提供了结构化的实验环境,在有限范围内开展合规探索与技术验证。西班牙政府允许地方监管部门与企业在真实场景中开展局部测试,对新技术进行可控范围的迭代验证。试错过程中,合规路径与实践经验不断积累,上级机构也得以识别技术风险与制度空白,为后续法规优化提供依据。测试机制协助监管机构在技术推进的早期阶段识别制度空白与风险瓶颈,进而为法规的动态修正提供支持。

(3)监督与反馈

监督与反馈是确保政策目标持续优化与动态调整的核心机制。该环节要求执行主体定期向监管机构提交阶段性报告,涵盖执行进度、实际成效、问题及改进建议,实现信息的上下流动,增强治理过程的透明性与响应性。这一机制有助于及时发现政策实施中的偏差与瓶颈,为后续的策略修正与制度完善提供实证支撑。通过持续反馈、经验归纳与横向比较,政策制定者基于现实运行状况修正框架目标与实施路径。

以新加坡生成式人工智能评估沙盒为例,参与企业需定期提交涵盖 AI 系统开发、测试数据及合规性审查的阶段报告,监管机构动态掌握技术风险与应用实效,优化监管工具与政策。沙盒机制通过结构化评估提升了模型的稳健性和合规性,确保创新在制度框架内有序发展。这一监督评估过程不仅实现了政策的实时调整,还增强了技术应用的责任基础和治理韧性。

(4)评议与调整

调整监管规范是实现政策持续优化的关键环节。随着基层执行过程中反馈和评估结果的积累,监管机构需根据实际问题动态调整现有监管框架和政策,以确保政策适应技术发展、市场需求和社会环境的变化。在人工智能监管沙盒中,参与者通过小规模测试和试点,能够识别初期规则在实际操作中的局

限性。基于这些反馈,监管机构审视现有法规,修改或引入新措施,确保技术创新在可控范围内推进[28]。

通过对实验主义治理的理论梳理与实践机制分析,本文提炼出实验主义治理在人工智能监管沙盒中的演进路径,具体表现为“框架目标设定—实验自主执行—监督与反馈—评议与调整”四个连续而互动的关键环节。论文将在下文中以 OECD 人工智能政策观察数据库中明确提出人工智能监管沙盒政策的九个国家为案例,围绕上述四个维度展开比较研究,剖析不同国家在沙盒治理中的制度设计差异与治理取向。

3 人工智能监管沙盒模式的国际比较

3.1 人工智能领域的监管沙盒需求

监管沙盒能够有效应对以人工智能为代表的新兴技术所带来的不确定性。根据 AI Act 的要求,成员国应在国家层面至少建立一个人工智能监管沙盒,并在 2026 年 8 月 2 日之前投入运行。尽管许多学者认为欧盟在人工智能治理方面相对严格,但监管沙盒已成为欧盟在确保安全的同时促进人工智能发展的关键工具,通过在受控环境中测试和验证应用,有效平衡了监管与创新之间的关系[29]。本文通过回顾“OECD 人工智能政策观察站”中来自 69 个国家的超过 1000 份 AI 相关政策,发现截至 2025 年 3 月,已有至少 9 个国家或地区正式建立了 AI 监管沙盒制度并公开相关制度文件。考虑到监管沙盒的设立需要具备良好的产业基础和成熟的市场环境,这一数量反映了监管沙盒作为促进人工智能健康发展的有效工具的快速演进。

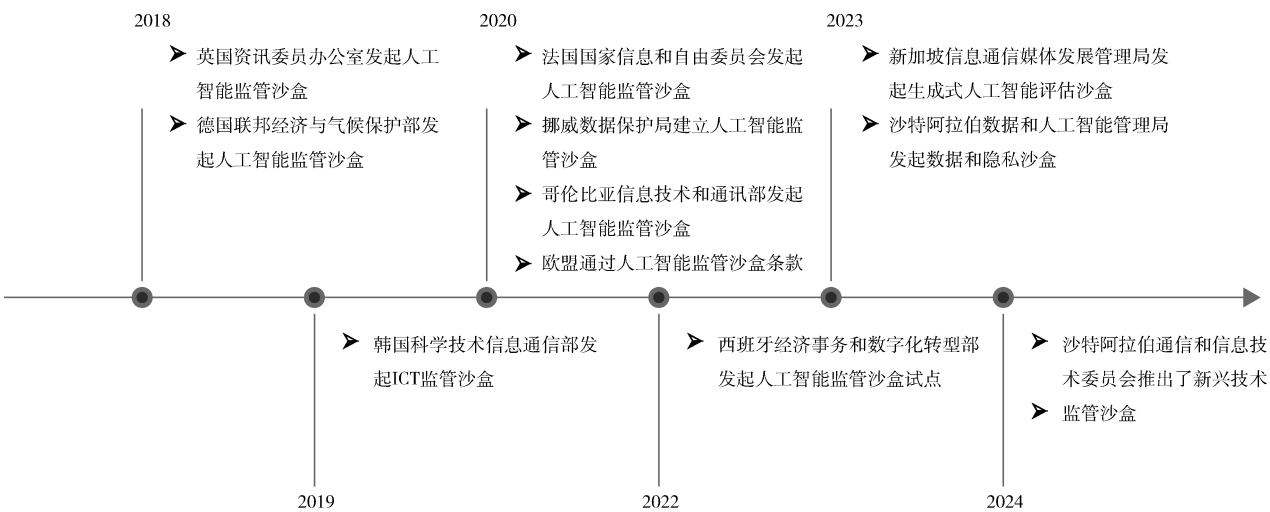

图 2 不同国家人工智能监管沙盒设立历程

Figure 2 History of the establishment of AI regulatory sandboxes in different countries

3.2 人工智能监管沙盒框架目标机制比较

不同国家的人工智能治理呈现多元化演进趋势。吴建南等[29]分析中、美、欧三方的人工智能安全治理政策文本,发现美国强调技术主导并注重安全,形成“以发展促安全”的战略;欧盟以伦理规范和风险控制为核心,体现“以安全促发展”的逻辑;中国采用“统筹发展与安全”的路径,力求在推动技术创新的同时建立韧性的安全治理体系。根据 OECD 人工智能政策观察站的政策文本,当前人工智能监管沙盒的治理取向分为三种模式:技术推进模式、风险控制模式和协同平衡模式。

(1)技术推进模式:以德国和韩国为代表的“技术推进模式”将监管沙盒作为推动新兴技术落地与商业转化的制度支点,聚焦技术验证、场景试运行和商业路径探索,通过小规模受控试验,为高风险技术提供可验证的发展空间,具有高度开放性与容错性。德国将沙盒纳入创新政策工具箱,平衡监管灵活性与公众信任,构建有序试验与制度演化的互动机制。

(2)风险控制模式:以英国、挪威、西班牙和哥伦比亚为代表的“风险控制模式”将监管沙盒作为识别前置性风险与嵌入伦理的工具,强调在创新初期确立责任边界和数据治理规范,通过制度化试验与政策预演,为人工智能发展奠定稳健的治理基础。英国推动“隐私内嵌设计”,在系统架构设计初期即

融入隐私保护要求,以前置性方式降低数据处理过程中的合规风险。

(3)协同平衡模式:协同平衡模式以新加坡、法国和沙特阿拉伯为代表,强调在技术创新与风险防控之间实现制度协同与动态调适。监管沙盒既提供制度弹性支持技术试验,又嵌入合规机制保障公共利益,平衡治理灵活性与规范性。法国监管沙盒以欧盟通用数据保护条例为基础,推动数据隐私法在人工智能数据处理中的应用。新加坡通过人工智能可信评估框架与测试工具箱提升透明度与可信度,推动可信产品验证与标准制定,引导技术发展。

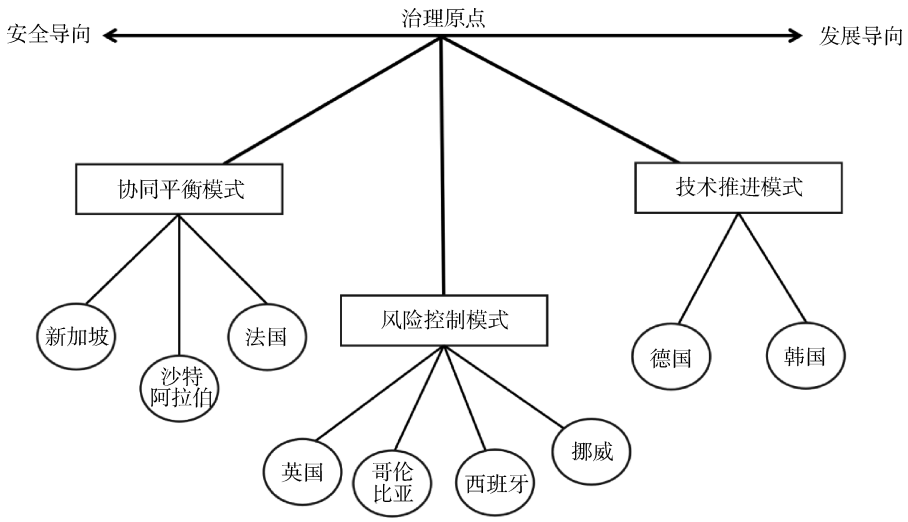

图 3　人工智能监管沙盒框架目标比较

Figure 3　Comparison of the objectives of AI regulatory sandbox frameworks

3.3　人工智能监管沙盒实验自主执行机制比较

人工智能监管沙盒的实验自主执行阶段是体现创新灵活性与监管平衡的关键环节。企业在监管框架下独立开展 AI 技术测试,但需遵守数据保护、算法透明性等规则。

(1)实验准备

实验准备通常会建立严格的审核标准和程序,确保进入监管沙盒的项目风险可控并具备创新潜力。前置审核流程需要提交项目方案、风险评估、合规性说明和伦理审查等材料。通过设立门槛,监管机构可以综合评估项目的合法性、安全性、社会影响和可行性。以 ICO 所采用的评估体系为例,申请项目必须围绕多个维度提交定性与定量支持材料。评估内容涵盖项目的创新程度、对公众利益的潜在影响、测试机制的可行性,以及是否具备有效的风险识别与缓释能力。项目需要说明在现有方法之上的创新突破,明确其所能带来的公共价值,通过合理的机制设计展现其落地实施的可操作性与稳健性。

(2)实验类型

按照治理对象的不同,沙盒实验实践可划分为三类:技术应用导向型监管、模型治理导向型监管、规则嵌入导向型监管。三类机制分别体现了人工智能在产业嵌入、系统责任与数据合规等不同层面上的制度回应路径。①技术应用导向型监管侧重于 AI 在关键行业的部署与合规,覆盖医疗健康、城市物流、公共服务等高风险领域。如英国“AI Airlock”项目专注于 AI 医疗设备测试,挪威测试患者数据授权机制,德国通过“真实实验室”推动城市自动化物流沙盒,沙特将智慧城市与物联网方案纳入沙盒支持范围。②模型治理导向型监管关注生成式人工智能、基础模型与通用系统等 AI 核心对象,强调系统责任边界与模型稳定性。西班牙在《人工智能法案》框架下测试高风险系统的操作性义务,新加坡通过 AI Verify 沙盒开展对大型语言模型的可解释性与安全性验证,挪威测试生成式 AI 在政务中的数据可信性与使用责任。③规则嵌入导向型监管聚焦 AI 系统中的数据合规、透明性与伦理保障,强调将制度规范前置嵌入技术设计之中。法国推广“隐私嵌入式设计”原则,西班牙强化合规文档与伦理审查机制,哥伦比亚通过预防性沙盒实现“默认合规”路径。

表 2 英国 ICO 监管沙盒入盒评估指标

Table 2 Assessment metrics for entry into the UK ICO regulatory sandbox

指标	1	2	3	4
创新指标	产品或服务仅为现有方法的改良,缺乏显著创新性	展现出与现有方法的差异,具备初步创新成果	创新成果有明确证据支撑,并与现有方法形成对比	有力证据显示创新具颠覆潜力,可能深刻影响行业
公共利益	产品仅利于自身,公众利益体现不足	产品可能带来公众收益,有初步量化与证据支持	公众利益明显,且有充分定性与定量证据支撑	产品具有实质性公众益处,证据充分且具变革性
计划可行性	沙盒设计偏离建议方法 时间表与资源需求不符 缺乏有效风险评估 退出机制存在缺陷	沙盒机制符合预期范围 时间表与资源匹配 风险评估较完善 退出计划清晰明确	机制设计合理,符合预期 部分风险已被识别评估 明确的退出计划	机制设计合理,符合预期 风险缓解与控制高效 明确的退出计划

资料来源:Information Commissioner's Office. Sandbox assessment criteria indicators[EB/OL]. https://ico.org.uk/media2/migrated/2618128/sandbox-criteria-indicators.pdf. 2025-04-01.

(3)实验流程

各国人工智能监管沙盒的政策目标与运行机制各不相同,但在实验流程设计上通常遵循"申请遴选—测试准备—受控实验—退出评估"的基本架构。在申请与遴选阶段,项目需提交技术方案、风险识别、数据处理与伦理说明等材料。英国和挪威采用滚动申请机制,增强沙盒的开放性;西班牙引入伦理审查,关注社会可接受性。测试准备阶段强调前置合规设计与责任机制,监管机构通过预沟通会议明确项目边界、时间表、数据规则及应急安排。德国通过"真实实验室"机制提供财政支持,强化沙盒的政策功能。受控试验阶段,多数国家采取"监管陪伴式"监督,动态调整测试路径。法国和哥伦比亚设有联合评估机制,确保伦理与法律同步介入,防止出现原本设定的时间限制、适用范围或风险边界逐步模糊化、放松甚至被规避的现象(即"沙盒漂移"现象)[30]。退出与评估阶段不仅包括总结报告,还涉及制度反馈与政策学习,德国和法国将此阶段输出用于法律优化,实现从实验机制向制度演化的闭环治理。

3.4 人工智能监管沙盒监督与反馈机制比较

实验主义治理常被视为一种在不确定环境下实现制度适应与政策演化的治理范式。沙盒运行强调的是过程中的持续反馈、灵活调整与制度进化,这种受控监督下"反馈—微调"的动态循环并不是其附属功能,而构成了其制度设计的本体性内容。各国人工智能监管沙盒的发展路径表明,监管并非在规则既定之后才启动评估与调整,而是将监督与反馈机制前置并深度嵌入整个运行过程,形成了"以监管为实验、以反馈促优化"的机制共识。

人工智能监管沙盒在监督与调整阶段主要体现出三种机制。(1)实时监测与反馈机制:在沙盒测试过程中,监管机构会嵌入实时监测机制,针对技术运行状态、伦理风险、合规问题等关键环节进行持续性的动态观察,收集和分析实时数据反馈。这一机制使监管能够紧密跟踪测试进程,及时发现风险与问题,为后续监管措施的动态调整奠定基础。(2)阶段性评估与即时调整机制:为了确保沙盒监管的响应性和适应性,一些国家(如德国、新加坡等),在沙盒运行过程中预设明确的中期评估节点,对监管要求进行及时的调整和微调。这种"阶段性反馈—即时修正"的方式提高监管的精准性与有效性。(3)裁量授权与规则弹性机制:部分国家在沙盒制度设计中采取了裁量授权型的路径,赋予监管机构较大的自主裁量空间,使其能够针对具体情况对规则进行解释、适用或豁免。这种制度设计在保证监管底线的同时,为监管机构提供了灵活的调整余地,有效应对人工智能技术发展中固有的不确定性,构建了具有弹性和容忍度的监管边界。

3.5 人工智能监管沙盒评议与调整机制比较

评议与调整机制侧重于将监管过程中积累的反馈信息系统性地转化为制度优化与政策完善的具体措施,形成政策持续演化与制度吸纳的有效路径。(1)规则与政策的调整机制:沙盒测试结束后,监管机构通常会依据实验数据,对法律条文、政策要求及技术标准进行系统性审视与修订。(2)产品与服务的优化机制:企业在沙盒试验中所获得的反馈信息不仅关注合规风险,也包括对技术逻辑、风险控制策略及数据处理方式的详细评估。企业可据此动态优化产品或服务,以提升市场竞争力和技术合规性。

(3)跨机构协作与标准建设机制:跨机构协作与标准建设是沙盒治理成功的关键。新加坡 AI Verify 项目整合模型开发者、企业、第三方测试机构及标准化组织,德国在推动人工智能沙盒实践中,构建涵盖多主体协调机制,逐步形成涵盖法律规范、信息服务和资金支持的综合性制度支撑体系,为沙盒治理的高效运行与标准建设提供了有力保障。

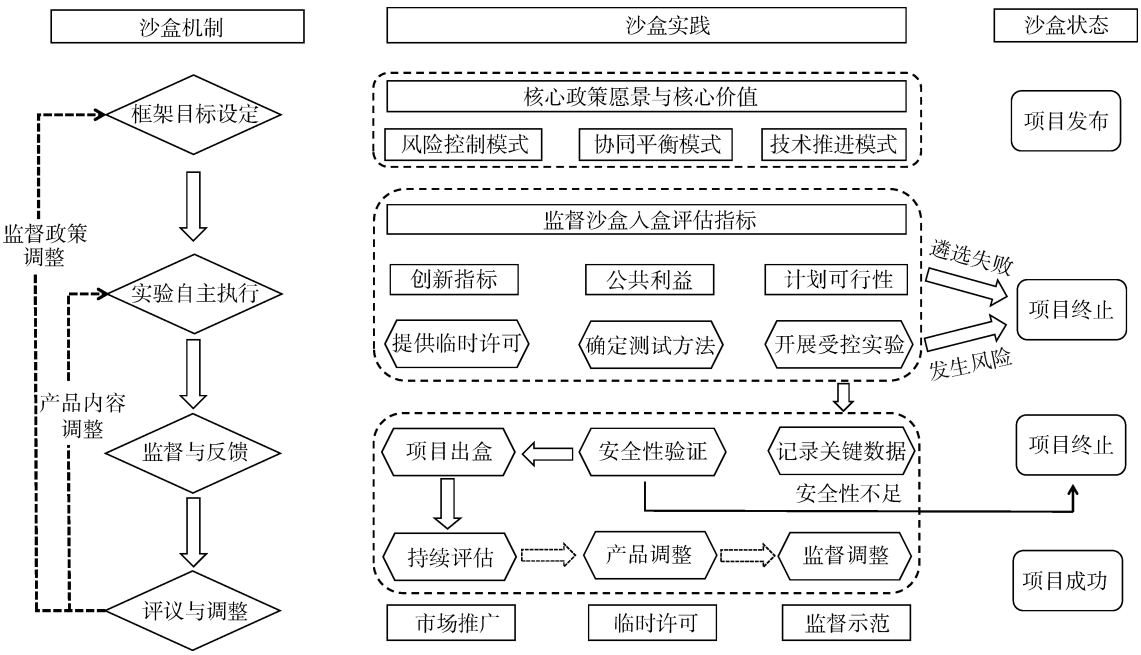

图 4　人工智能监管沙盒运行流程

Figure 4　AI regulatory sandbox operation flow

4　未来监管展望

我国人工智能监管沙盒仍处于探索阶段,人工智能监管沙盒的制度建设重点应围绕分领域布局、可量化的申请与评估标准、多元主体协同机制以及国际规则协调路径等方面展开系统探索,以构建更具弹性、透明性与适应性的 AI 监管体系。

4.1　推进分领域监管沙盒的体系化布局

人工智能监管沙盒的制度配置呈现出明显的领域差异性,不同国家在技术应用、模型治理与规则嵌入等方面各有侧重。这种“按需设箱”的治理逻辑启示我国在未来制度设计中,应突破通用性监管的路径依赖,推动人工智能监管沙盒在重点行业、核心技术与敏感数据场景中分类设立与差异治理。一方面,针对医疗健康、城市治理、制造业、教育等关键行业,设立专题沙盒,重点围绕技术可行性、伦理风险和数据流通开展受控测试;另一方面,关注基础模型、训练数据、生成式 AI、跨模态系统等新兴形态的系统性风险,推动设立以模型为对象的专属监管机制,形成“数据—模型—应用”三级协同的治理架构。通过分领域布局,增强制度弹性和风险识别能力,探索中国场景的监管样态。

4.2　建立可量化的申请标准与评分机制

建立科学、清晰且可操作的申请标准体系,将是提升沙盒运行质量与治理效能的关键。借鉴英国 ICO 实践,我国可探索建立“四维一体”评分体系,即围绕创新性、公共价值、风险可控性与机制合理性设定具体评估指标与分级赋分标准。创新性可依据技术突破与行业颠覆潜力分级评估,公共价值可量化其服务范围、受益群体与社会影响广度,风险控制关注数据合规、伦理预案与应急能力。评分机制贯穿项目的申请、测试与退出全过程,形成“动态评估—反馈调整”的闭环结构。同时,建议配套公布评分规则与评估,提升机制透明度与可预期性,从而为具备创新潜力和公共价值的人工智能项目提供受控试验空间,增强激励功能。

在制度实现路径上,推动中央层面出台统一的监管沙盒申请规范文件,地方与行业在框架下进行灵活适配,形成“统一规则—差异嵌套”的协同格

局。通过标准体系建设,提高监管效率、保障程序正义,引导企业形成可预期的技术合规路径,营造良好的创新生态与制度信任基础。

4.3 构建协同共治的合作机制架构

人工智能监管沙盒有效运行依赖规则设计的科学性与多元治理主体之间的高效协同。联合评估机制通过引入伦理专家与行业代表参与测试反馈与合规评判,可以构建跨主体的协同监督路径,新加坡依托 AI Verify 项目便是典型代表。该项目通过政府、企业、科研机构和社会组织的共同参与,形成“企业测试—政府引导—技术反馈”的制度闭环。

我国可在既有试点基础上,推动构建“沙盒合作平台”,联合政府部门、高校、科技企业与第三方机构,形成制度化、多层级的常设合作网络,共同参与项目筛选、风险评估、伦理审查与政策反馈。同时,鼓励行业协会、地方政府与标准化组织参与,在中央统筹下实现“上下联动、横向协同”的运行机制。配套制定协作指引、责任清单与流程守则,将有助于明确各方权责边界,防止协作失衡与制度冗余。

4.4 探索跨国互认与标准协调机制

全球人工智能治理尚处于多轨并行的早期阶段,不同国家在沙盒框架、评估标准与合规要求上存在较大差异,容易导致规则碎片化与“监管套利”现象,阻碍全球协同治理的形成。推动跨国监管沙盒的互认与合作试点,将成为提升制度兼容性与促进国际信任的关键路径。欧盟 AI Act 提出沙盒制度的区域内适用框架,西班牙已通过 AI 沙盒先行探索高风险系统与基础模型的监管操作化,形成可供参考的跨境治理范式。

我国应积极参与全球 AI 治理标准的制定与沙盒机制的互认对接,推动在亚太区域和“一带一路”框架下建设人工智能监管沙盒合作网络。可考虑建立“中外联合监管试验区”,在数据合规、模型责任、伦理规范等方面开展协同测试,提升制度互通性和标准操作性。构建跨国沙盒合作平台,开展标准对话与成果互认证,提升我国在全球人工智能治理中的规则塑造能力与制度话语权。通过推进分领域机制建设、确立科学量化的申请标准、完善多元协同的合作平台以及参与全球沙盒规则的互认协作,监管沙盒将不仅是风险控制的容器,更是制度创新的孵化器,为中国式人工智能治理现代化提供重要的政策工具和制度支撑。

参考文献:

[1] 贾开,蒋余浩. 人工智能治理的三个基本问题:技术逻辑、风险挑战与公共政策选择[J]. 中国行政管理, 2017(10): 40 – 45. Jia K, Jiang Y H. Three fundamental issues in the governance of artificial intelligence: Technical logic, risk challenges, and public policy choices [J]. Chinese Public Administration, 2017(10): 40 – 45.

[2] 李洋,薛澜. 颠覆、调适与协同:责任伦理视域下生成式人工智能的多主体治理机制研究[J]. 电子政务, 2025(7):40 – 49. Li Y, Xue L. Disruption, adaptation and synergy: Research on the multi – agent governance mechanism of generative artificial intelligence from the perspective of responsibility ethics [J]. E – Government, 2025(7): 40 – 49.

[3] 胡滨. 金融科技、监管沙盒与体制创新:不完全契约视角[J]. 经济研究, 2022, 57(6): 137 – 153. Hu B. Fintech, regulatory sandbox and institutional innovation: From the perspective of incomplete contracts[J]. Economic Research Journal, 2022, 57(6): 137 – 153.

[4] 陈振其. 监管沙盒:数据要素治理新方案[J]. 图书馆论坛, 2024, 44(4): 138 – 147. Chen Z Q. Regulatory sandbox: A new solution for data element governance[J]. Library Tribune, 2024, 44(4): 138 – 147.

[5] 王频,陈云良. 数字经济时代金融消费者保护新范式——基于监管沙盒模式的展开[J]. 科学决策, 2023(1): 119 – 132. Wang P, Chen Y L. A new paradigm for the protection of financial consumers in the digital economy era – Based on the development of the regulatory sandbox model [J]. Scientific Decision – Making, 2023(1): 119 – 132.

[6] 杨丰一. 人工智能“监管沙盒”:理论溯源、实践考察与制度构建[J]. 中国行政管理, 2024, 40(12): 25 – 35, 56. Yang F Y. "Regulatory sandbox" for artificial intelligence: Theoretical origin, practical investigation and institutional construction[J]. Chinese Public Administration, 2024, 40(12): 25 – 35, 56.

[7] 张红. 监管沙盒及与我国行政法体系的兼容[J]. 浙江学刊, 2018(1): 77 – 86. Zhang H. Regulatory sandbox and its compatibility with China's administrative law system[J]. Zhejiang Academic Journal, 2018(1): 77 – 86.

[8] Fahy L A. Fostering regulator – innovator collaboration at the frontline: A case study of the UK´s regulatory sandbox for fintech[J]. Law & Policy, 2022, 44(2): 162 – 184.

[9] Alaassar A, Mention A L, Aas T H. Exploring how so-

cial interactions influence regulators and innovators: The case of regulatory sandboxes[J]. Technological Forecasting and Social Change, 2020, 160: 120257.

[10] Bromberg L, Godwin A, Ramsay I. Fintech sandboxes: Achieving a balance between regulation and innovation [J]. Journal of Banking and Finance Law and Practice, 2017, 28(4): 314 - 336.

[11] Truby J, et al. A sandbox approach to regulating high - risk artificial intelligence applications [J]. European Journal of Risk Regulation, 2022, 13(2): 270 - 294.

[12] 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. Xue L, Zhao J. Moving towards agile governance: Exploring the development and regulatory models of emerging industries[J]. Chinese Public Administration, 2019(8): 28 - 34.

[13] 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. Jiang L D,Xue L. The Challenges of the times and the paradigm change in the governance of China's new - generation artificial intelligence[J]. Journal of Public Management, 2022,19(2): 1 - 11,164.

[14] 郭小东. 大模型时代全球人工智能治理的挑战与中国方案[J]. 科学学研究, 2025,43 (4):694 - 702. Guo X D. Challenges and Chinese solutions for global artificial intelligence governance in the era of large - scale models[J]. Studies in Science of Science, 2025, 43 (4):694 - 702.

[15] 薛澜,贾开,赵静.人工智能敏捷治理实践:分类监管思路与政策工具箱构建[J]. 中国行政管理, 2024, 40(3): 99 - 110. Xue L, Jia K, Zhao J. The practice of agile governance of artificial intelligence: The classification supervision idea and the construction of the policy toolkit[J]. Chinese Public Administration, 2024, 40 (3): 99 - 110.

[16] 邓悦,许弘楷,王诗菲. 人工智能风险治理:模式、工具与策略[J]. 改革, 2024(1): 144 - 158. Deng Y, Xu H K, Wang S F. Artificial intelligence risk governance: Models, tools and strategies[J]. Reform, 2024 (1): 144 - 158.

[17] 何宇华,李霞. 生成式人工智能虚假信息治理的新挑战及应对策略——基于敏捷治理的视角[J]. 治理研究, 2024, 40(4): 142 - 156,160. He Y H, Li X. New challenges and coping strategies for the governance of false information in generative artificial intelligence - from the perspective of agile governance[J]. Governance Research, 2024,40(4): 142 - 156,160.

[18] 武振国. 欧盟人工智能的实验主义治理路径及中国借鉴[J]. 西北大学学报(哲学社会科学版), 2024, 54(6): 153 - 164. Wu Z G. The experimentalist governance path of EU Artificial Intelligence and its reference for China[J]. Journal of Northwest University (Philosophy and Social Sciences Edition), 2024, 54 (6): 153 - 164.

[19] 苏竣. 开展人工智能社会实验探索智能社会治理中国道路[J]. 中国行政管理, 2021(12): 21 - 22. Su J. Conducting social experiments on artificial intelligence to explore the Chinese path of intelligent social governance[J]. Chinese Public Administration, 2021(12): 21 - 22.

[20] 曾雄,梁正,张辉. 人工智能软法治理的优化进路:由软法先行到软法与硬法协同[J]. 电子政务, 2024 (6): 96 - 107. Zeng X,Liang Z, Zhang H. The optimization path of soft law governance of artificial intelligence: From Soft Law first to the coordination of soft law and hard Law[J]. E - Government, 2024(6): 96 - 107.

[21] 曹建峰. 迈向负责任 AI:中国 AI 治理趋势与展望[J]. 上海师范大学学报(哲学社会科学版), 2023, 52(4): 5 - 15. Cao J F. Moving towards responsible AI: Trends and prospects of AI governance in China [J]. Journal of Shanghai Normal University (Philosophy & Social Sciences Edition), 2023,52(4): 5 - 15.

[22] Sabel C F, Zeitlin J. Learning from difference: The new architecture of experimentalist governance in the EU [J]. European Law Journal, 2008, 14(3): 271 - 327.

[23] 陈靖,洪伟. 试验还是实验? 试点与实验主义治理的比较[J]. 科学学研究, 2020, 38(9): 1537 - 1544. Chen J, HongW. Experiment or test? A comparison between pilots and experimentalist governance[J]. Studies in Science of Science, 2020,38(9): 1537 - 1544.

[24] 刘太刚,邓正阳. 实验主义治理:公共治理的一个新路径[J]. 北京行政学院学报, 2020(1): 34 - 42. Liu T G,Deng Z Y. Experimentalist governance: A new path for public governance[J]. Journal of Beijing Administrative College, 2020(1): 34 - 42.

[25] 康镇. 政策试点的实验主义治理逻辑与转型进路[J]. 求实, 2020(4): 56 - 69,111. Kang Z. The experimentalist governance logic and transformation approach of policy pilots[J]. Seeking Truth, 2020(4): 56 - 69,111.

[26] Kumm M. Constitutionalism and experimentalist governance[J]. Regulation & Governance, 2012, 6(3): 401 - 409.

[27] Sabel C F, Zeitlin J. Experimentalist Governance in The European Union: Towards ANew Architecture[M]. Oxford University Press, 2010.

[28] Khanal S, Zhang H, Taeihagh A. Building an AI ecosystem in a small nation: Lessons from Singapore's journey to the forefront of AI[J]. Humanities and Social Sciences Communications, 2024, 11(1): 1 - 12.

[29] 吴建南,马太平,周磊. 人工智能安全治理的规则体系:国际比较与协同进路[J]. 中国行政管理, 2024, 40(12): 6 - 14. Wu J N, Ma T P, Zhou L. The rule system for the security governance of artificial intelligence: International comparison and the path to coordination[J]. Chinese Public Administration, 2024, 40(12): 6 - 14.

[30] Truby J, Brown C, Dahdal A. A sandbox approach to regulating high - risk artificial intelligence applications[J]. European Journal of Risk Regulation, 2022, 13(2): 270 - 294.

From static to agile: Governance mechanism of artificial intelligence regulatory sandboxes

LYU Yue[1,3], CHEN Xu[1], PENG Zi - xuan[2], XU Yu[1]

(1. School of International and Public Affaris, Shanghai Jiaotong University, Shanghai 200030, China;
2. School of Public Health, Southeast University, NanJing 210009, China;
3. Shanghai University of Engineering Science, Shanghai 201600, China)

Abstract: Traditional regulatory models have become increasingly inadequate in addressing the complexity and rapid evolution of artificial intelligence (AI) technologies. Regulatory sandboxes, which originated in the fintech sector, have been expanded into the AI domain, emerging as tools to balance security and development. These sandboxes have been adopted by numerous countries worldwide. At present, the design of the AI regulatory mechanism remains in the exploratory stage, with divergent practices across countries in key aspects such as legal authorization, regulatory boundaries, and governance models. This paper reviews the origin, evolutionary path and international application of AI regulatory sandboxes, and adopts the framework of "experimentalist governance" to guide the construction and operation of regulatory sandboxes. We summarize the operation processes of artificial intelligence regulatory sandboxes in 9 countries around the world, and extract the mechanisms of framework goal setting, independent experimental execution, supervision and feedback, evaluation and adjustment. This mechanism illustrates the applicability of regulatory sandboxes in enabling dynamic policy adjustment, fostering multi - stakeholder collaboration, and promoting regulatory innovation. It facilitates the continuous optimization of policy objectives by encouraging effective interactions among regulators, innovators, and consumers through real - time feedback and coordinated efforts, thereby fostering a positive relationship between technological innovation and regulatory compliance. The development of AI regulatory sandboxes should focus on sector - specific regulatory approaches, the formulation of measurable application criteria, and the enhancement of international cooperation and mutual recognition mechanisms to address global governance challenges posed by AI technologies. By promoting the localized evolution of sandbox systems and actively participating in the formulation of international standards, China can enhance its AI governance capacity cross domestic and international regulatory systems.

Key words: technology governance; regulatory sandbox; governance reform; experimentalist governance; comparative study