人工智能审计的理论内涵、问责边界与框架构建
——基于负责任创新视角

崔永梅,杨婷羽,应文池,曹润泽

(北京交通大学 经济管理学院,北京 100044)

[摘　要]人工智能(AI)的快速发展在推动社会进步的同时也引发了诸多风险,而人工智能审计(AI 审计)作为一种新兴审计范式,能够推动 AI 负责任创新。遵循 AI 审计"是什么""审什么""怎么审"的逻辑思路,在明确 AI 审计理论内涵的基础上,探索 AI 审计的问责边界,梳理审计主体、原则和实施路径,从而构建 AI 审计框架。研究发现,第一,AI 审计意味着"审 AI",是对 AI 负责任创新的问责机制;第二,AI 审计聚焦 AI 的可负责性和可审计性,明确 AI 创新链各环节的审计风险和问责边界,形成以 AI 模型为核心,向内拓展至数据、算法,向外拓展至产品和生态系统的审计范围;第三,内部审计、社会审计与国家审计各自发挥独特的作用,在技术可靠、安全合规和伦理道德目标下,贯穿审计计划、实施到完成的全过程。通过结合 AI 负责任创新视角与审计问责机制,构建一个本土化、时代性的 AI 审计框架,推动中国审计学自主知识体系构建。

[关键词]人工智能审计(AI 审计);人工智能(AI);负责任创新;审计框架;审计问责

[中图分类号]F239.34　　[文献标志码]A　　[文章编号]1004-4833(2025)04-0011-12

一、引言

近年来,人工智能(Artificial Intelligence,AI)发展迅猛,在自然语言生成、图像合成、个性化推荐、自动化招聘等领域得到广泛应用(如 ChatGPT、DeepSeek、文心一言、Kimi、豆包、Midjourney 等),显现出强大的创造力和高度的自主性,在推动行业智能化变革、提高组织效率方面成效显著[1]。然而,AI 在推动社会进步的同时,也引发了数据滥用、算法歧视、隐私侵犯和社会偏见等新型风险[2-4],对公共利益和社会伦理构成挑战[5]。特别是 AI 在决策过程中的不透明性和不可解释性,即背后隐藏的技术"黑箱",进一步加剧了 AI 风险治理的复杂性。

面对上述挑战,如何推动 AI 负责任创新(Responsible Innovation),使 AI"向善、为善"逐渐成为研究焦点[6]。负责任创新强调技术进步需与伦理、法律和社会责任并行,以实现可持续发展[7]。在这一背景下,各国纷纷呼吁 AI 治理,发展负责任、可信赖的 AI。我国《新一代人工智能治理原则——发展负责任的人工智能》便提倡将责任内嵌于 AI 技术创新议程,确保其透明性和公平性[8]。尽管当前 AI 治理实践已聚焦技术改进和公法规制,但仅靠原则难以有效约束开发者和运营者对 AI 的行为负责[9]。换句话说,现有的 AI 监管体系忽略了第三方问责的重要性[10],在开拓新范式、融合多路径方面仍显不足[11]。那么,当 AI 产生风险时,如何追溯并明确责任归属?又如何建立一种审查机制以保障其负责任创新?

为解决上述问题,人工智能审计(以下简称"AI 审计")应运而生。作为一种独立、客观、创新型的审计机制,AI 审计能够对 AI 全生命周期创新链条中的技术可靠性、安全合规性和伦理道德进行全面评估[8,10],从而对 AI 与利益相关者的要求和期望的匹配程度作出合理的审计判断。正如财务报表审计评价财务报表是否公允反映改变了 20 世纪企业的运营方式,AI 审计将通过获取和分析 AI 风险相关的审计证据,并将审计结果传达给利益相关方,有效推动 AI 负责任创新[11]。AI 审计的优势在于技术、法律和伦理三元共治,不仅能为技术层面的透明性、可靠性与可解释性提供鉴证,还能对法律层面的数据偏见、隐私泄露进行评价,甚至关注社会层面的伦理

[收稿日期]2024-12-03

[基金项目]国家社会科学基金青年项目(24CJY130);中国国家铁路集团有限公司项目(2023F011)

[作者简介]崔永梅(1969—),女,山东烟台人,北京交通大学经济管理学院教授,博士生导师,博士,从事审计、内部控制与风险管理研究;杨婷羽(1998—),女,山东潍坊人,北京交通大学经济管理学院博士研究生,从事审计、技术创新与风险管理研究,通信作者,E-mail:tyyang@bjtu.edu.cn;应文池(1981—),男,江西南昌人,北京交通大学经济管理学院副教授,硕士生导师,博士,从事新兴数字技术应用与管理研究;曹润泽(2000—),男,山东济宁人,北京交通大学经济管理学院博士研究生,从事审计理论与实践研究。

道德和责任表现。可见,审计是完善 AI 风险治理和责任追溯体系的重要一环,在发展负责任 AI 的多元监管中发挥独特作用。

尽管审计机制在风险应对领域已有成熟的应用,如 IT 审计、模型审计、算法审计,但仍然难以适应 AI 发展态势[12]。孙甲奎指出,在人工智能等新技术的冲击下,审计的运行环境、组织方式、目标对象等都发生了深刻变化,传统审计理论和方法已不完全适用[13]。与传统审计中注册会计师面临的固有风险、控制风险相比,衍生风险的叠加使 AI 审计风险呈现出新的特点。与算法审计相比,AI 的技术复杂性和社会影响力更高;而与模型审计相比,AI 涵盖更广泛的数据、产品乃至生态系统。这些差异对传统审计提出了新的要求。为加快构建中国审计学自主知识体系,发展中国特色社会主义审计理论,需要推动审计学科交叉融合,立足中国实际,解决中国问题[14],将传统审计框架针对 AI 特性进行创新。

目前,针对 AI 审计的研究虽已初步展开,但尚未形成系统的理论框架,尤其是在 AI 审计“是什么”“审什么”“怎么审”方面,尚缺乏一套相对成熟、科学的体系[10]。为此,本文提出一个致力于推动 AI 负责任创新的审计框架,核心内容包括:第一,回答 AI 审计“是什么”的问题,明确 AI 审计的理论内涵,包括含义、本质、特征和功能定位;第二,回答 AI 审计“审什么”的问题,基于对 AI 创新链的阶段划分,分析其在各个环节的衍生风险,从而明确 AI 审计的问责边界;第三,回答 AI 审计“怎么审”的问题,形成政府主导、社会协同、企业支持的中国特色 AI 审计主体,遵循目标导向、依法合规、独立客观原则,从事前、事中和事后对传统审计路径进行拓展,由此涵盖审计主体、审计原则与审计实施路径。本文将审计机制与 AI 负责任创新相结合,构建适应中国情境的 AI 审计框架,以期丰富 AI 治理领域的审计理论,从审计角度促进 AI 安全、向善,推动本土化、时代性的中国审计学自主知识体系建设。

二、AI 负责任创新与审计问责:文献综述与经验借鉴

(一)文献综述

1. AI 风险与负责任创新

AI 的快速迭代和广泛应用引发了多重风险[15-16]:一是技术层面的风险,现代 AI 的复杂性和自主性增加了其在决策过程中的不可预测性和不可解释性[12];二是法律层面的风险,数据泄露、隐私保护不足与安全漏洞频现,促使关于 AI 法律监管的讨论[17];三是社会层面的风险,AI 生成的内容可能包含潜在的刻板印象,导致对特定群体的歧视性内容传播,加深社会偏见和不平等现象。更严重的是,三种风险可能互相交织,如 AI 内部“黑箱”的隐秘性和复杂性,进一步增加了社会公众识别和感知其潜在偏见的难度,从而对弱势群体构成了更大的风险[16]。

在此背景下,负责任创新(Responsible Innovation,RI)逐渐成为全球科技治理的核心议题。负责任创新理念最早起源于欧盟“第六框架计划”和“第七框架计划”,其核心宗旨在于确保科技进步在推动社会发展的同时,不会对伦理、社会和环境造成负面影响。其特征包括三方面:一是伦理导向,要求创新过程中注重公平性和包容性,避免算法歧视与社会偏见的扩大;二是透明与可解释性,增强公众对技术的信任感;三是社会适应性与合规性,即在遵循法律法规的基础上,确保技术对公共利益的正向贡献[7,18]。值得强调的是,负责任创新并非“反创新”,而是一种适当的制衡,旨在确保 AI 在造福社会的同时有效防范风险[19]。欧盟在“地平线 2020”规划中进一步强调了负责任创新,提出了负责任创新的六个关键要素,包括公众参与、伦理、科学教育、性别平等、开放获取和治理,倡导在技术研发的各个阶段纳入社会参与、伦理考量和透明监督,以避免创新伴随着的潜在风险。

在负责任创新导向下,各国纷纷制定相关的 AI 治理框架。2019 年 4 月,欧盟发布了《可信赖的人工智能伦理准则》,明确了 AI 应满足合法、合乎伦理和稳健三大要求,从而推动“负责任 AI”的发展[12]。同样,我国也在逐步构建自主的 AI 治理框架。2023 年 7 月,我国国家互联网信息办公室发布了《生成式人工智能服务管理暂行办法》,强调在 AI 技术应用中优先选择安全可信的技术手段,并明确了责任主体的问责机制。新一代人工智能发展规划推进办公室成立了国家新一代人工智能治理专业委员会,发布了《新一代人工智能治理原则——发展负责任的人工智能》,提出人工智能发展应遵循安全可控、尊重隐私、敏捷治理、共担责任等八项原则,为推动 AI 的负责任创新奠定了制度基础。

2. 负责任 AI 的审计问责机制

当前关于 AI 负责任创新的研究多基于技术视角,将负责任创新视为企业内部的“闭门”合规问题,忽视了多元利益相关者的参与及非技术层面的治理需求[20]。美国纽约大学 AI Now 研究所发布的《年度 AI 现状报告》指出,仅靠技术解决 AI 风险问题存在很大局限性,建立有效的 AI 外部问责机制日益迫切。负责任创新的核心是确保 AI 在技术开发和社会应用过程中,符合社会标准、伦理道德和法律法规,而 AI 审计正是实现这一目标的重要手段。AI 审计起源于算法审计,最初用于检测算法应用中可能产生的歧视、偏见及对个体和组织的不良影响。在过去十年间,审计作为一种治理工具,从对算法权力进行有效监督,逐步扩展为对 AI 进行系统化审查[12]。AI 审计不仅需要关注技术可靠性,还应纳入社会伦理考量,以确保 AI 系统的公平和正义。

相比审计的鉴证、评价等功能,现有研究更多地将 AI 审计视为独立的问责机制。2021 年,联合国教科文组织发布的《人工智能伦理问题建议书》强调,审计能够确保在全生命周期内对 AI 及其影响实施问责。AI 审计问责机制主要由可负责性和可审计性两个部分组成[11]。一方面,可负责性可以具体化为时间、信息和行动三个维度,强调在 AI 全生命周期中建立一条清晰的责任链,使系统开发者和使用者能够解释 AI 决策的合理性。相关方有权要求披露决策背后的原因,以确保任何不当行为都可以追溯到特定行为者[21]。例如,荷兰已经为政府使用 AI 制定了一个审计框架,侧重于谁承担什么责任以及责任体现在哪里。另一方面,可审计性则侧重于外部审查和独立验证的能力,在 AI 审计过程中,审计师不依赖 AI 开发和运行人员的解释,而是通过审计程序获取客观的审计证据,识别和评估 AI 在实际应用中的潜在风险,对 AI 是否按预期的方式工作(如技术安全可靠)或符合适用标准(如符合伦理道德、法律法规)做出客观评价[12],并针对审计发现提出改进建议。尤为重要的是,有效的 AI 审计能够实现可负责性和可审计性的融合,不仅能够识别技术、法律和伦理等方面的错报,还可以通过问责机制,明确谁对 AI 的决策和行为负有责任,使各方利益相关者对审计结果负责[22],从而推动 AI 负责任创新。

3. AI 审计的内容研究

关于 AI 审计内容的研究大多分散在审计主体、审计方法与审计效果方面。在 AI 审计主体方面,前期大部分研究集中于内部审计,因其与外部审计相比具有更高的访问权限,如果设计得当,可以实现问责制,但也可能因利益关系而导致虚假保证。例如,Google 内部 AI 团队人员因发表对大规模语言模型的批评意见而被解雇,Facebook 员工因提出偏见问题而被内部压制[23]。因此,引入独立的第三方审计对 AI 问责至关重要,其有效性取决于清晰的审计范围、审计师独立性与胜任能力、充分的数据访问权限及完善的后续披露机制[24]。此外,还可以将 AI 审计嵌入国家专项审计或经济责任审计,如在政策跟踪审计中根据 AI 相关政策,审查被审计单位的落实情况[12]。

在 AI 审计方法方面,大量研究集中于评估 AI 公平性与可解释性的工具。Costanza-Chock 等对 152 名审计师进行了调查,发现 62% 的审计师使用现有工具,如 AI Fairness 360、Scikit Fairness 或 Parity,但只有 7% 为其整体审计协议使用了标准化框架[24]。在 AI 审计后果方面,Raji 等指出 AI 审计结果可以为 AI 通用标准制定提供信息[10]。由于缺乏对性能和证据的具体期望,许多关于人工智能伦理和问责制的讨论都受到了阻碍[9]。而审计机制的一大优势在于可以帮助揭示技术背后的具体问题,例如,IEEE 自动面部分析技术行业标准便受到性别审计的影响。有研究呼吁未来的 AI 所有者和运营商需根据明确的标准进行独立审计,强制披露审计结果,推动 AI 审计师的评估和认证[24]。

(二)国际经验与中国实践探索

1. AI 审计的国际经验

全球范围内,各国和国际组织在 AI 审计领域进行了积极探索。美国政府问责署发布了 AI 问责框架,美国国防部发布了《负责任的人工智能指南》。2023 年,美国国家标准与技术研究院发布《人工智能风险管理框架》,建议由独立第三方或非系统一线开发人员的专家对 AI 进行评估。纽约市通过立法要求使用 AI 进行自动化招聘决策的公司需接受独立审计。欧盟《通用数据保护条例》引入内部审计机制,《数字服务法案》引入算法问责和透明度审计的条款,要求超大型在线平台每年向具备独立性的外部审计提交报告。此外,欧盟《人工智能法案》首次提出实施风险分级监管,要求对高风险 AI 系统进行内部合规审计,并对生物识别等敏感技术实施外部审计。英国审计署对政府 AI 模型管理运行情况进行了全面审计,并基于调查结果提出 AI 审计框架。基于这一框架,英国审计署审计了英国财政部、预算责任办公室等 17 个中央政府部门关键模型风险,并提出了审计建议。

国际组织层面,国际内部审计师协会早在2017年便发布了AI审计框架,包括网络弹性、人工智能能力、数据质量、数据架构和基础设施、衡量绩效、伦理学和黑匣子七个审计要素。

2. AI审计的中国探索

中国结合本土国情,已在AI负责任治理方面展开了初步探索,但有待引入完善的AI审计机制。2023年,国家互联网信息办公室等七部门发布了《生成式人工智能服务管理暂行办法》,要求AI服务提供者进行安全评估。《新一代人工智能治理原则——发展负责任的人工智能》提出"共担责任""敏捷治理"等八项原则,强调技术可控、隐私保护和伦理合规。《中华人民共和国数据安全法》和《中华人民共和国个人信息保护法》为AI审计提供了法律依据,要求企业确保数据采集、存储和使用的合规性。2025年,抖音公布将建设安全与信任中心网站和线下公示展厅,推进算法和平台治理透明化。毕马威中国开发了"AI治理成熟度评估模型",通过建立AI审计专门的知识库、制度库、审计模型和方法论,从企业战略发展、数据使用、建模、评估、部署监控五大领域,诊断企业AI应用的透明度和可问责性。

(三)研究述评

现有研究表明,审计问责在推动AI负责任创新的过程中扮演着关键角色,但是仍存在研究缺口。第一,现有文献大多聚焦算法透明性或模型可解释性的局部技术审计,而对数据、产品乃至AI生态系统的审计却鲜有涉及。事实上,从数据收集、模型训练到产品部署,每个环节都可能潜藏数据泄露、算法偏见、模型黑箱等风险。因此,未来研究需要关注AI技术创新链条全生命周期的审计,尤其关注安全合规、伦理道德等关键领域,以推动AI负责任创新。第二,审计作为保障AI负责任创新的核心机制,当前仍存在定位模糊的问题。AI技术链条涉及数据提供者、算法开发者、模型训练者以及用户、监管机构、社会组织等多方利益相关者,各方在审计过程中的责任划分尚不明晰。因此,AI审计应超越单纯的技术审查和鉴证,明确审计作为一种问责机制的具体内容和标准。第三,当前AI审计相关文献呈现出碎片化特征,缺乏中国特色本土化、时代性的审计框架。已有研究对AI审计主体的讨论主要分散在审计独立性、审计方法的选择以及审计后果的评估,且AI审计经验主要以欧美国家为主,中国AI审计发展较为缓慢。如何将不同审计范围、主体、原则、实施路径等中国特色审计知识纳入一个系统的AI审计框架,探索中国自主AI审计理论和方法,是当前亟待解决的问题。

三、AI审计理论内涵

为回答AI审计"是什么"的问题,本文基于对现有研究的系统梳理,提出AI审计的核心内涵,包括AI审计的含义、特征及功能定位。

(一)AI审计的含义、本质及特征

相比目前较多文献探究如何在审计中利用AI技术,即"用AI辅助审计"[25-26],本文研究的AI审计聚焦对AI风险进行审计鉴证、评价与问责,即对AI执行审计。在此定义下,AI审计的本质在于通过获取和分析审计证据,揭示AI风险并提供审计建议。类似于财务报表审计评价财务报表是否公允反映,在AI开发、传播和应用全过程中,审计师结合技术监督、法律保障与伦理评价,判断AI是否达到预期标准、是否符合利益相关者的期望,传达审计结果,以促进AI负责任创新,增强社会信任。

由此可见,在构建中国审计学自主知识体系的背景下,AI审计具有区别于传统审计的核心特征:一是跨学科融合性,AI审计涉及审计学、计算机科学、伦理学和法律等多学科交叉领域,要求审计人员具备多维度的专业知识;二是全生命周期覆盖,从数据开发、算法设计、模型搭建到产品部署及生态影响,AI审计贯穿AI创新链的全过程,确保审计结果的全面性与系统性;三是独立性与客观性,AI审计通过第三方独立评估,保障审计结果不受利益相关方的干扰。

(二)AI审计的功能定位

不同于传统审计,AI审计不仅能够发挥鉴证与评价功能,还是一种对负责任创新的问责机制。其核心目标是通过审计问责,确保AI在技术可靠性、法律合规性和伦理一致性等方面达到预期标准,以增强利益相关者对AI在可靠、安全、道德等方面的信赖程度。

具体而言,AI审计强调明确的审计判断和独立的评估过程,以衡量AI部署的实际表现是否符合社会和法律的期望。为了使这些审计判断更加具体,需要在AI部署的现实与期望之间进行审计判断,以将透明性、公平

性或安全性与预期期望联系起来。需要注意的是,由于招聘、医疗保健、刑事司法或社会服务交付等特殊领域的特定审计判断,它们对于不同 AI 所表达的预期期望也有所不同。例如,有些 AI 专门为技术可靠性进行审计,而另一些 AI 则更侧重法律合规性或伦理一致性期望。因此,通过 AI 审计问责机制,可以评估和诊断对 AI 负责任创新的一系列合规、伦理问题以及更广泛的社会不公平,而不仅限于 AI 技术的安全可靠。

四、AI 审计风险、范围与问责边界

在 AI 审计中,“审什么”是一个根本性问题,其核心取决于审计师对 AI 能否负责任创新的了解程度。AI 的复杂结构和“黑箱”特性,使其缺乏内部可见性,审计师对 AI 系统内部逻辑的了解往往存在知识鸿沟,容易依赖 AI 开发人员的解释,无法做到充分、独立的评估,进而导致决策逻辑出现错误归因和不确定性。为弥合知识鸿沟,需实现业审融合,需要做到以下两点:一是深入理解 AI 创新链条及其固有风险衍生;二是在识别 AI 固有风险的基础上,设定 AI 审计边界。

(一)AI 可负责性:AI 创新链与审计风险衍生

AI 审计具有跨学科、跨阶段的特征。为实现 AI 的负责任创新,应系统审视整个 AI 技术创新链,而非仅关注孤立的算法、模型等风险。AI 技术创新链涵盖 6 个环节,包括基础研发、模型开发、内容生成、平台部署、场景渗透与用户触达(图 1)。链条中的每个阶段都可能衍生出不同类型、相互交织的审计固有风险,包括自身缺陷或使用不当,从而对 AI 的可负责性提出严峻挑战。因此,明晰 AI 创新链及风险衍生是 AI 可审计性的前提。

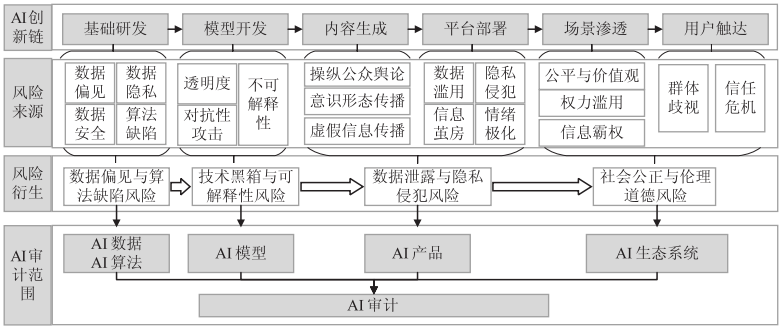

图 1　AI 技术创新链与风险衍生

1. 基础研发阶段:数据偏见与算法缺陷风险

在基础研发阶段,AI 的核心构件(数据集和算法)决定了系统整体性能。若数据质量不佳或训练数据集的代表性不足,则易在算法中引入系统性偏见。例如,大型语言模型的训练数据如果包含性别、种族等方面的隐性偏见,那么这些偏见将不可避免地反映在模型的输出中[27]。此阶段的典型风险主要包括两个方面:一是数据偏见,训练数据的不均衡或选择性过滤会导致 AI 系统在处理特定用户或群体时表现出歧视性行为;二是算法缺陷,算法设计中嵌入的主观性或商业逻辑导向可能导致偏见决策。这些偏见不仅损害用户体验,还可能违反公平性和伦理标准[22]。因此,基础研发阶段需引入 AI 数据审计和 AI 算法审计,以识别潜在偏见源。

2. 模型开发阶段:技术黑箱与可解释性风险

AI 模型开发与训练阶段涉及复杂的算法调试和优化,是技术实现的核心环节。然而,深度学习模型的高度复杂性和“黑箱”性质使得外部观察者难以理解其决策逻辑,增加了 AI 应用的风险,尤其在医疗、法律等高风险领域更是如此。这种风险一方面体现为模型透明度不足,缺乏透明度的 AI 模型会限制用户对其决策过程的理解,从而影响信任度和接受度[27]。另一方面,未经审计的 AI 模型更容易受到对抗性攻击,攻击者通过输入特定数据误导模型,导致不准确甚至危险的决策输出。因此,提升 AI 模型的可解释性,并加强对“黑箱”问题的应对,是 AI 模型审计中必须关注的重点。

3. 内容生成与平台部署阶段:数据泄露与隐私侵犯风险

在这一阶段,AI 产品需采集分析用户数据并根据需求生成内容。然而,AI 生成的内容往往涉及信息传递、用户数据处理及隐私保护等多个层面的风险。例如虚假信息传播风险,AI 生成内容可能被恶意利用来操纵公众舆论或传播虚假失真信息,导致社会信任危机[3]。此外,平台在收集和使用用户数据时,若未能采取有效的隐私保护措施,将极易导致用户隐私泄露。生成内容还可能未经授权使用受版权保护的素材,进而引发法律纠纷。基于上述风险,这一阶段的 AI 产品审计应着重评估用户数据管理及隐私合规,以确保其输出符合法律法规标准。

4. 场景渗透与用户触达阶段:社会公正与伦理道德风险

在 AI 系统场景渗透以及最终用户触达阶段,其决策结果会对社会行为和伦理道德产生深远影响。例如,基

于 AI 技术的决策系统在刑事司法、医疗诊断和招聘等领域的应用,可能会对社会公平和价值观构成威胁。一方面,需防范社会偏见与歧视风险。AI 决策可能基于不公平的数据或算法,导致对特定群体的歧视。二是 AI 权力的滥用。AI 系统的广泛应用可能导致 AI 权力的无意识扩散,引发 AI 对社会行为和公共决策的隐性控制。因此,这一阶段的 AI 系统审计应当引入社会公正和伦理道德的评估,以确保 AI 的可持续和负责任应用。

(二)AI 可审计性:审计范围与问责边界

AI 审计的"审什么"不仅包括对 AI 风险的评估,还需明确其在不同审计范围下的问责边界。本文定义了五个主要审计范围,并结合具体对象阐述每一范围内的问责内容。

1. AI 审计范围

以往关于 AI 审计的研究主要将算法作为审计的主要对象,而没有保留被审计实体范围的开放性[21]。基于 AI 创新链的复杂性,为了有效应对上述不同阶段可能衍生的潜在风险,本文将 AI 审计的范围定义为广义的五个层级,即以 AI 模型审计为核心,向内涵盖数据和算法审计,向外延伸至产品和生态系统的全覆盖审计(图 2)。第一层级是 AI 数据审计。AI 的基础是大规模的数据集,确保数据集的充分性、适当性和代表性至关重要。AI 数据审计旨在识别数据偏差以降低偏见风险。第二层级是 AI 算法审计。AI 算法审计关注 AI 模型的训练过程和算法结构,分析算法的透明度和鲁棒性,以确保其输出结果的准确性和公平性。第三层级是 AI 模型审计。AI 模型审计旨在揭示 AI 的"黑箱"特性,提升其可解释性和透明度,以便更好地评估其在实际应用中的表现。第四层级是 AI 产品审计。对于嵌入 AI 模型的实际产品,AI 产品审计关注其技术性能、法律兼容性、用户隐私保护及伦理合规。第五层级是 AI 生态系统审计。AI 的影响已从单一产品延伸至整个社会技术系统,因此有必要从生态系统视角审视 AI 的社会影响力。AI 生态系统审计关注 AI 的社会交互及其整体影响,旨在建立跨系统的审计问责机制。

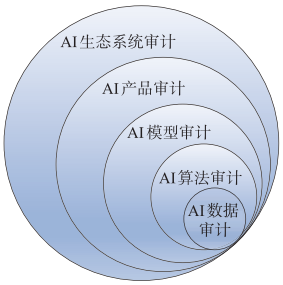

图 2　AI 审计研究范围

2. AI 审计问责边界

负责任创新要求明确不同 AI 审计范围内需满足的技术标准,如数据的收集、处理和存储需要符合隐私规范。这能为审计人员在进行审计活动时提供判断依据,确保审计人员能够依据统一的技术规范进行操作,避免因标准不明而导致的审计偏差或不一致性。此外,还应界定各类行为应由哪些责任主体承担。表 1 明确了 AI 审计的问责边界,为不同审计范围的 AI 界定了责任主体。

表 1　不同审计范围下的 AI 审计问责内容

审计范围	技术标准	审计问责边界	责任主体	实践案例
AI 数据	数据的收集、处理和存储符合数据治理标准 数据完整性、数据泄漏防护以及对数据偏见的识别 数据隐私保护机制的设立	数据来源 数据质量 数据使用	数据提供商 数据处理方	若训练数据中男性相关术语占据过多比例,审计需指出数据偏见并提出改进建议
AI 算法	算法决策溯源 算法的解释性和可解释性 对算法结果的公平性和偏见的控制	算法透明度 算法公正性	算法开发者	某些推荐算法对特定人群存在优先性或排斥性,AI 审计需明确算法开发者对决策偏见的纠正责任
AI 模型	模型验证、测试、优化、维护、监测等过程性问题模型的过拟合、欠拟合等问题 模型在不同环境下的适用性 模型在现实场景中的表现偏差 模型安全漏洞和对抗性攻击的防护	模型稳健性 模型安全性	模型训练者	对自动驾驶 AI 模型进行审计时,审计人员可通过模拟不同光照条件或交通标志变化,评估其对外界扰动的敏感性及应对机制
AI 产品	产品生命周期管理、产品升级和迭代过程中风险的管理 产品采购、部署、使用、管理控制涉及人员的胜任力 产品成本收益分析	功能有效性 产品合规性	产品开发者	在电商平台使用的推荐系统中,需审查其是否通过不当手段收集用户数据,导致用户信息过度暴露或隐私泄露风险
AI 生态系统	长期质量鉴证 用户的互动 信息披露 对社会公正、就业潜在影响评估 环境福祉考虑 多方协作过程中责任的划分	伦理一致性 社会经济影响 公平与价值观 绿色可持续性	系统集成商、多方利益相关者(用户、监管机构、社会组织)	某智能招聘系统的应用是否导致对特定性别或种族的歧视,加剧不公平现象;AI 技术的能源消耗及其碳排放影响,并明确多方利益相关者的责任划分

设定审计边界的目的还在于界定 AI 在具体应用中的责任归属,同时认识到审计的局限性,明确审计至多提供合理保证而非绝对保证,边界之外不适合审计。例如,在 AI 数据审计中,审计师只能对已知的数据集和使用

场景进行评估。此外,由于 AI 可能带来严重但渐进的社会风险,审计只是更广泛 AI 问责的一个组成要素,因此,在设定审计问责边界时,应考虑可审计性,不能将 AI 审计视为解决 AI 弊病的唯一手段。

五、AI 审计主体、原则与实施路径

基于对 AI 创新链、审计风险与审计问责边界的探索,结合国内外对 AI 审计的理论研究以及现实实践,本文从 AI 审计主体、审计原则与审计实施路径三个方面回答 AI 审计“怎么审”的问题。

(一)AI 审计主体

AI 审计主体具有多元性,根据审计特征、内容和作用的差异,可以分为内部审计、社会审计和国家审计三类。

1. 内部审计

内部审计主要由企业内部的审计部门或独立审计团队进行,旨在确保企业在 AI 技术开发与应用中的内部控制,测试 AI 是否符合公司或社会期望。其优势在于访问权限广、介入时点早[22],能深度介入全过程,包括数据集选择、算法内部决策逻辑、模型训练等核心环节。

然而,内部审计通常存在利益冲突和独立性问题。内部审计师由企业选聘并支付报酬,因而受到组织关系的制约,当 AI 开发由高管直接推动时,审计师可能因压力而难以独立发表审计意见。这可能导致审计报告偏向企业利益,而非向外部利益相关者客观披露风险。例如,若纠正 AI 偏见可能削弱 AI 性能,内部审计便可能选择忽视,Twitter、Facebook 都曾因对公平问题短视,反而将注意力转移到隐私保护、错误信息而受到批判。因此,在内部审计中要求审计师具备严谨性和职业胜任能力,并引入外部监督机制,建立更为独立的审计委员会成为增强其公信力的有效措施。

2. 社会审计

社会审计通常由第三方机构(如会计师事务所)、非营利组织或学术机构受托进行,其审计目标通常是为了维护社会公众利益,向外部利益相关者发出 AI 公平性、透明性和合规性等信号[10]。相比其他审计主体,社会审计独立性高、公开性强,具有更大的灵活性和自主性,能够最大限度地减少利益冲突和不当影响,通过公开披露审计结果,达到以外部压力促进 AI 负责任创新的目的。

然而,社会审计也存在专业胜任能力不足、信息获取难、审计资源有限等问题。由于 AI 审计尚处于初级阶段,专业审计人员和审计机构比较匮乏,除了一些技术类咨询公司或者毕马威等审计咨询类公司涉及 AI 审计外,大部分社会审计机构更多呈现出“参与式审计”的特点。此外,由于社会审计机构通常与被审计对象无直接利益关系,获取关键数据和访问权限存在较大困难,常依赖公开数据和用户反馈,限制了审计的深度和准确性,影响审计质量[10]。因此,社会审计机构需要与政府和企业建立更紧密的合作关系,以提高其在 AI 审计中的数据可获得性和影响力。

3. 国家审计

国家审计由政府部门(如审计署)实施,重点关注公共资产和国有资金中的 AI 应用。随着数字化转型不断深化,AI 已经嵌入政务服务、公共安全和社会治理等核心领域,因此,国家审计对 AI 的监管和问责显得尤为重要。国家审计的优势在于权威性强、强制性高,可以通过立法和政策制定,确保 AI 在公共领域的安全、透明和公平应用。

国家审计面临的挑战在于时效性和针对性。AI 发展迅速、复杂多变,国家审计难以在不受到干扰的情况下进行深入的审查,容易滞后于技术发展,或因过度干预导致政策不稳定或企业抵触。因此,对于国家审计,可以将 AI 审计嵌入现有的信息系统审计(IT 审计)框架中,通过一般控制审计和应用控制审计相结合,全面评估被审计单位的 AI 风险。还可以将 AI 审计嵌入专项审计项目,如在政策跟踪审计中,以国家出台的《新一代人工智能伦理规范》等 AI 相关制度为依据,审查政策的落实情况;或是在领导干部经济责任审计中审查领导干部对于 AI 风险防控的责任。此外,还可以将影响国家安全、存在重大风险的 AI 作为国家专项审计项目进行单独立项[12]。

内部审计、社会审计和国家审计作为 AI 审计的三大主体,具备各自独特的审计特点(表 2)。在负责任创新的视角下,三者应形成有效的协同问责机制,成为政府主导、社会协同、企业支持的中国特色 AI 审计主体。内部

审计侧重于企业对 AI 的内部控制,社会审计强调 AI 的外部信任,而国家审计则聚焦国有资金范围内 AI 的公共利益。此外,为应对审计人员与 AI 开发人员之间的技术鸿沟,提升审计人员的专业能力,审计机构应着重提升审计人员在 AI 技术、数据分析和伦理评估方面的专业能力。在审计过程中,聘请计算机科学、数据科学等领域的技术专家,或组建跨学科的审计团队,包括审计专家、技术专家和伦理专家。审计机构还可以开发专门的 AI 审计工具,帮助审计人员更高效地识别和评估 AI 审计风险。

表 2　不同 AI 审计主体及其特点

审计主体	审计需求	优势	不足	提升措施	案例
内部审计	优化企业对于 AI 内部控制,降低风险	访问权限广、介入时点早	利益冲突、独立性弱	加强技术培训、引入外部专家	Google、IBM、Meta 等通过内部审计,利用 AI 360 Toolkit、Fairness Flow 等工具,检测 AI 偏见和歧视
社会审计	增强对于 AI 的社会信任,强化外部监督	独立性高、公开性强	专业胜任能力不足、信息获取难、审计资源有限	开发 AI 审计工具、建立跨学科团队	毕马威开发了“毕马威治理成熟度评估模型”,建立专门的知识库、制度库、审计模型和方法论,从企业战略发展、数据使用、建模、评估、部署监控方面对 AI 执行审计
国家审计	保障 AI 相关的公共利益、推动政策落实	权威性强、强制性高	时效性低、针对性弱	加强政策支持、推动 AI 审计标准化	英国审计署对财政部、预算办公室 AI 模型管理运行情况进行全面审计

(二)AI 审计原则

结合 AI 审计内涵、范围与边界,负责任创新视角下的 AI 审计应遵循目标导向、依法依规与独立客观三大原则。

1. 目标导向原则

无论通过何种主体进行审计,审计范围、重点和结论都将取决于其目标。借鉴欧盟人工智能高级专家组发布的《可信人工智能道德准则》和我国国家新一代人工智能治理委员会发布的《新一代人工智能治理原则——发展负责任的人工智能》的要求,结合 AI 的复杂性、应用场景的多样性,我们认为 AI 审计可以根据目标差异分为三类:一是基于技术可靠目标,旨在问责 AI 数据、算法与模型是否符合行业标准与技术规范,审计内容包括模型源代码安全检测、算法性能测试、数据清洗与标注规范性等;二是基于安全合规目标,关注 AI 产品的数据隐私保护与信息安全问题,内容包括数据加密、访问控制、隐私协议评估,以及用户数据的使用授权与脱敏处理,帮助用户在使用 AI、防控风险等方面做出知情选择;三是基于伦理道德目标的 AI 审计,聚焦 AI 系统在伦理道德层面是否符合社会规范,审计内容包括 AI 权力偏见检测以及自动化决策的公平性评估。

2. 依法合规原则

依法依规是 AI 审计的核心原则,要求审计行为在法律和政策框架内进行。在负责任创新视角下,AI 审计在遵循行业标准的同时,更应该严格遵守法律法规。不同审计目标对应的法规有所差异,例如,技术可靠审计需要遵循《互联网信息服务算法推荐管理规定》《互联网信息服务深度合成管理规定》等;安全合规审计需要依照《生成式人工智能服务管理暂行办法》《中华人民共和国数据安全法》《中华人民共和国网络安全法》《中华人民共和国个人信息保护法》等;伦理道德审计需要参考《新一代人工智能治理原则——发展负责任的人工智能》《国家新一代人工智能标准体系建设指南》《新一代人工智能伦理规范》等,确保 AI 系统符合伦理道德要求。总的来看,虽然我国现有法规已将 AI 纳入监管,但内容较为分散,给 AI 审计实践带来了挑战。借鉴欧盟的《人工智能法案》,推动制定专门的法律法规是中国特色 AI 审计未来的探索方向之一。

3. 独立客观原则

独立性和客观性是审计工作的基本保障,AI 审计亦不例外。在“审 AI”的过程中,审计独立性面临独特的挑战,特别是在复杂算法和模型审查中,被审计方对技术细节的掌控可能导致信息不对称,进而影响审计的客观性与公正性。为确保 AI 审计的独立性,对于外部审计而言,社会或国家审计机构与被审计方之间应无利益关系;对于内部审计而言,需保持对业务部门的独立监督。客观性体现在审计师对 AI 保持职业怀疑,确保审计报告的客观公正。在该原则要求下,审计师应独立评估被审计单位内部控制机制是否健全,独立分析 AI 决策逻辑,减少对被审计方解释的依赖。同时,引入审计轮换制度,遵守职业道德,以确保 AI 审计结果的公平公正。此外,AI 审计独立客观原则还意味着公开的信息披露,如定期向外界公布 AI 审计报告,明确说明所依据的审计证据、审计标准及相关审计过程,接受外部监督,增强公众信任。

基于对 AI 审计三大原则的探讨,表 3 展现了不同原则涉及的核心维度及具体内容。

表 3　AI 审计三大原则及其审计重点

审计原则	核心维度	审计内容	实践步骤
目标导向	技术可靠	审查 AI 数据、算法、模型的技术规范和行业标准，包括数据清洗、算法性能测试、源代码安全性检测等	明确审计目标→界定审计范围→分配资源→设计审计程序→形成审计报告
	安全合规	聚焦 AI 产品的数据隐私保护与信息安全，如数据加密、访问控制、隐私协议评估等	
	伦理道德	评估 AI 生态系统是否符合社会规范，例如权力偏见检测、自动化决策公平性审查等	
依法合规	以技术可靠目标为导向的 AI 审计	遵循《互联网信息服务算法推荐管理规定》《互联网信息服务深度合成管理规定》	收集相关法规→审计证据采集→审查合规性→形成审计建议
	以安全合规目标为导向的 AI 审计	遵循《生成式人工智能服务管理暂行办法》《中华人民共和国数据安全法》《中华人民共和国网络安全法》《中华人民共和国个人信息保护法》	
	以伦理道德目标为导向的 AI 审计	遵循《新一代人工智能治理原则——发展负责任的人工智能》《国家新一代人工智能标准体系建设指南》《新一代人工智能伦理规范》	
独立客观	独立性	对 AI 保持职业怀疑，独立分析 AI 决策逻辑，减少对被审计方依赖，引入审计轮换制度，遵守职业道德	审计主体隔离（利益冲突审查）→设计独立审计程序→定期披露结果
	客观性	定期公开披露 AI 审计报告，接受外部监督，增强公众信任	

（三）AI 审计实施路径

为将 AI 审计的技术标准落实至操作层面，需重点构建 AI 审计的工作标准，即“怎么审”。在负责任创新视角下，AI 审计的实施路径应遵循全生命周期审计框架，从审计证据的收集、分析到最终的审计结论生成，分为事前（审计计划）、事中（审计实施）及事后（审计报告）三个阶段（图 3）。需要指出的是，对于 AI 数据、算法、模型、产品以及生态系统五大审计范围而言，审计实施路径的差异主要体现在事中（审计实施）阶段，特别是实质性程序中。因此，本文在这一部分细化了对不同范围内的审计路径分析。

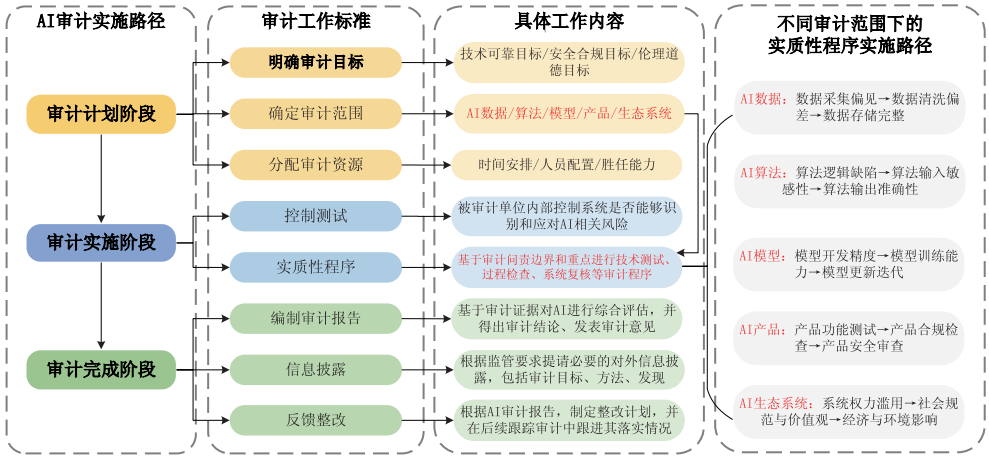

图 3　AI 审计实施路径

1. 事前：审计计划

AI 审计的计划阶段是整个审计过程的基础，其主要任务是明确审计目标、确定审计范围、分配审计资源，以及识别和评估潜在风险。首先，基于 AI 的审计需求和应用场景，明确审计的总体目标，包括技术可靠性、安全合规性和伦理道德规范等。其次，确定审计的具体范围，可以是 AI 数据、算法、模型、产品以及生态系统，也可以是范围之间的组合。再次，根据审计任务的规模和复杂程度，合理分配审计资源，包括时间安排、有经验的审计人员配置以及必要的技术支持。同时，应充分考虑 AI 审计过程中可能涉及的跨学科知识需求，确保团队具备相应的胜任能力。最后，识别并评估 AI 衍生风险，包括数据偏见与算法缺陷、技术“黑箱”与可解释性、数据泄露与隐私侵犯、社会公平与伦理道德等。

2. 事中：审计实施

在 AI 审计实施阶段，审计人员应采取控制测试和实质性程序相结合的方法，以有效应对已识别的 AI 风险。第一，控制测试主要评估被审计单位内部控制系统是否能够识别和应对 AI 相关风险，包括对公司内部监督流程、管理层的风险认识及其应对措施进行测试，以验证 AI 相关的内部控制有效性。第二，实质性程序主要依据前文探讨的审计问责边界和重点，基于不同审计范围实施技术测试、过程检查、系统复核等多种评估方法。为避免“一刀切”的模式，这一阶段应当针对数据、算法、模型、产品和生态系统的特性设计专属审计程序，实施差异化审计路径。

（1）AI 数据

AI 数据是所有 AI 整体系统的基础，其质量直接决定后续模型与算法的可靠性。AI 数据审计的主要目标是

验证 AI 数据在采集、清洗和存储方面的技术可靠性。审计实施过程中,在数据采集方面,应审查数据集中是否存在代表性偏差或历史偏见,变更记录是否透明且无篡改;在数据清洗阶段,需检测数据中的缺失值、异常值及潜在的偏差问题,同时审查数据清洗规则是否具备合理性与透明性;在数据存储方面,审计人员需确保存储系统的安全性和完整性,重点检查加密措施、备份机制及访问权限的设置是否符合技术标准。若发现错报,则根据问责机制追究数据提供商或处理方的责任。

(2)AI 算法

AI 算法容易因设计缺陷而导致决策不公,因此,AI 算法审计的主要目标是评估 AI 算法透明、公正方面的技术可靠性。为此,审计人员需从算法逻辑、测试角度展开工作。一方面,通过审查算法开发文档,验证其假设、逻辑规则和实现代码是否存在缺陷;另一方面,通过实施测试,分析 AI 对输入数据的敏感性、一致性及算法输出的准确性。若发现错报,则根据问责机制追究算法开发者的责任。

(3)AI 模型

AI 模型的技术"黑箱"属性与可解释性风险对审计深度的要求更高,因此,AI 模型审计的主要目标是评估 AI 模型开发、训练和部署的技术可靠性。在开发阶段,审计人员需检验模型的精度、召回率等关键指标,在训练阶段,通过量化测试模型的预测能力、鲁棒性及可解释性,检验模型的抗干扰能力和应急机制,在模型更新与迭代阶段,审计人员需对模型更新过程的记录机制进行核查,确保相关操作可追溯且在可控范围内。若发现错报,则根据问责机制追究模型训练者的责任。

(4)AI 产品

AI 产品作为最终输出直接面向用户,其信息泄露和隐私侵犯是核心风险点,因此 AI 产品审计的主要目标是评估 AI 产品功能的安全合规性。在功能测试环节,需评估不同使用环境下 AI 产品核心功能是否按设计正常运行,重点关注用户交互界面的安全性;在合规检查方面,结合行业监管标准和适用法规,评估产品的合规性。此外,还需审查产品是否可能引发隐私侵犯或数据泄露问题。若发现错报,则根据问责机制追究产品开发者的责任。

(5)AI 生态系统

AI 生态系统的复杂性和外部性问题容易引发广泛的社会风险,因此,AI 生态系统审计需要加以重点考量。一方面,应对生态系统内各模块及外部合作伙伴的协同机制进行审查,以确保不存在 AI 权力滥用问题。另一方面,需重点审查 AI 系统的外部性问题,如符合社会规范和价值观、产生的经济和环境影响,以保障生态系统的负责任创新。若发现错报,则根据问责机制追究系统集成商以及多方利益相关者的责任。

3. 事后:审计完成

在 AI 审计的事后阶段,审计人员需对审计发现进行总结和汇报,披露审计报告和可持续改进。基于收集到的审计证据,审计主体对 AI 进行综合评估并得出审计结论。在此过程中,首先需要与被审计单位治理层和管理层进行充分沟通,报告审计发现并就关键问题达成一致意见。其次,形成正式的审计报告后,还需根据监管要求,提请必要的对外信息披露,包括审计目标、方法、发现及建议,以保护内外部利益相关者。最后,建议被审计单位根据 AI 审计报告,制定整改计划,并在后续跟踪审计中跟进其落实情况,以促进 AI 负责任创新。

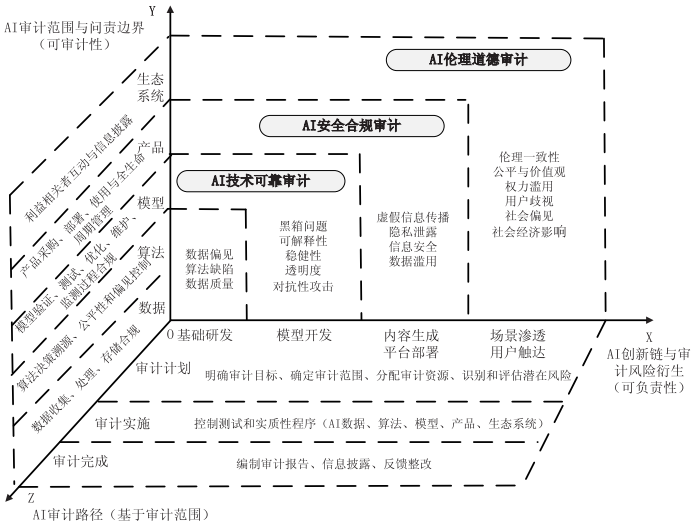

图 4　AI 审计框架

六、基于负责任创新视角的 AI 审计框架构建

国际上已出现 AI 审计相关的框架或规范,而国内审计框架尚未完全适应 AI 发展,亟须总结中国审计经验,制定适用于本土情况的中国特色 AI 审计框架。负责任创新视角下,本文明确了 AI 审计的理论内涵与问责边界,进而构建 AI 审计框架(图 4)。AI 审计框架通过矩阵形式,将三个维度有机结合,包括体现可负责性的 AI 创新链与审计风险衍生、体现可审计性的 AI 审计范

围与问责边界、基于审计范围的 AI 审计路径。

首先,从坐标轴来看:AI 创新链(X 轴)涵盖 AI 全生命周期,包括基础研发、模型开发、内容生成、平台部署、场景渗透与用户触达;AI 审计范围(Y 轴)可以划分为五个层次,包括 AI 数据、算法、模型、产品和生态系统全覆盖;框架明确了 AI 审计时间维度下的实施路径(Z 轴),包括审计计划阶段、审计实施阶段和审计完成阶段。其次,从象限来看,AI 审计框架旨在系统化识别 AI 创新链中各个环节特有的审计风险衍生(X 轴与 Y 轴围成的 XOY 部分),进而明确审计问责边界与重点(Y 轴与 Z 轴围成的 YOZ 部分),引导审计目标的选择(技术可靠、安全合规、伦理道德),基于审计实施路径中阶段的划分,为 AI 审计工作提供理论支撑和操作指引(X 轴与 Z 轴围成的 XOZ 部分)。

七、结语

当前,AI 的快速发展对其治理体系提出了新要求,而审计问责机制可以在其中发挥巨大作用。本文基于负责任创新视角,系统性地回答了 AI 审计“是什么”“审什么”与“怎么审”三大关键问题,有助于推动中国审计学自主知识体系构建,也为未来 AI 审计实践提供了理论支持。第一,关于 AI 审计“是什么”的问题,本文界定了 AI 审计的理论内涵,厘清了 AI 审计旨在“审 AI”的含义、本质、特征及功能定位。第二,关于 AI 审计“审什么”的问题,本文揭示了 AI 创新链各个环节衍生的审计风险,明确了 AI 负责任创新的技术标准,进而确定了 AI 审计范围和问责边界。第三,关于 AI 审计“怎么审”的问题,本文从审计主体、审计原则和审计实施路径三个方面探索了 AI 审计工作标准。

基于以上三大关键问题,本文构建了一个本土化、时代化的 AI 审计框架,有助于为政策制定者和内外部审计人员提供一种结构化的工具,以便更有效地识别、评估和应对 AI 审计风险,从而落实对 AI 负责任创新的问责机制。本文将 AI 负责任创新视角与审计问责机制相结合,立足中国实际,解决中国问题,有助于完善我国 AI 治理体系,拓展审计应用场景以及与其他学科融合创新,从而构建中国审计学自主知识体系。

本文仍存在一些不足之处:研究结论主要基于理论分析、文献综述与实践总结,未来研究可以通过案例研究等方法,检验框架的有效性。此外,随着 AI 迅速迭代升级,AI 审计的适用范围、审计目标等内容也在动态变化中,未来研究需持续关注 AI 技术发展和政策动向,将 AI 审计融入现有审计准则体系中,甚至探索制定专门的 AI 审计准则,以适应 AI 时代的挑战。

参考文献:

[1]白晓红,季瑞华. AI 对审计工作深度影响研究——基于 AI 对注册会计师行业审计影响与应对的视角[J]. 审计研究,2024(4):56-64.

[2]Mökander J,Schuett J,Kirk H R,et al. Auditing language models:A three-layered approach[J]. AI And Ethics,2024,4(4):1085-1115.

[3]Benjamin R. Assessing risk,automating racism[J]. Science,2019,366(6464):421-422.

[4]肖红军. 算法责任:理论证成、全景画像与治理范式[J]. 管理世界,2022(4):200-226.

[5]高锦萍,白羽新,高居平,等. 人工智能时代的会计伦理:内涵、转向与考量[J]. 会计研究,2022(3):17-27.

[6]周旅军,吕鹏. “向善”且“为善”:人工智能时代的算法治理与社会科学的源头参与[J]. 求索,2022(1):135-142.

[7]Burget M,Bardone E,Pedaste M. Definitions and conceptual dimensions of responsible research and innovation:A literature review[J]. Science and engineering ethics,2017,23:1-19.

[8]林慧涓,陈宋生. 构建多维度 AI 审计框架思考[J]. 会计之友,2023(23):32-37.

[9]Mittelstadt B. Principles alone cannot guarantee ethical AI[J]. Nature machine intelligence,2019,1(11):501-507.

[10]Raji I D,Xu P,Honigsberg C,et al. Outsider oversight:Designing a third party audit ecosystem for ai governance[C]. Proceedings of the 2022 AAAI/ACM Conference on AI,Ethics,and Society,2022.

[11]Birhane A,Steed R,Ojewale V,et al. AI auditing:The broken bus on the road to AI accountability[C]. 2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML). IEEE,2024.

[12]王玉凤. 模型算法审计:理论内涵、国际经验与审计框架[J]. 审计研究,2023(3):11-18.

[13]孙甲奎. 面向中国式现代化的审计学科功能定位与创新发展[J]. 南京审计大学学报,2025(1):1-11.

[14]秦荣生. 以习近平总书记关于审计工作的重要论述为指引建构中国审计学自主知识体系[J]. 审计研究,2024(5):17-25.

[15]陈雄桑. 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(10):138-141.

[16]唐要家,唐春晖. 基于风险的人工智能监管治理[J]. 社会科学辑刊,2022(1):114-124.

[17]周辉. 人工智能基础模型安全风险的平台治理[J]. 财经法学,2024(5):3-22.
[18]李娜,陈君. 负责任创新框架下的人工智能伦理问题研究[J]. 科技管理研究,2020(6):258-264.
[19]McGregor L,Murray D,Ng V. International human rights law as a framework for algorithmic accountability[J]. International & Comparative Law Quarterly,2019,68(2):309-343.
[20]Krafft P M,Young M,Katell M,et al. An action-oriented AI policy toolkit for technology audits by community advocates and activists[C]. Proceedings of the 2021 ACM conference on fairness,accountability,and transparency,2021.
[21]Loi M,Spielkamp M. Towards accountability in the use of artificial intelligence for public administrations[C]. Proceedings of the 2021 AAAI/ACM Conference on AI,Ethics, and Society,2021.
[22]Raji I D,Buolamwini J. Actionable auditing revisited: Investigating the impact of publicly naming biased performance results of commercial ai products [J]. Communications of the ACM,2022,66(1):101-108.
[23]Ebell C,Baeza-Yates R,Benjamins R,et al. Towards intellectual freedom in an AI Ethics Global Community[J]. AI and Ethics,2021,1:131-138.
[24]Costanza-Chock S,Raji I D,Buolamwini J. Who audits the auditors? Recommendations from a field scan of the algorithmic auditing ecosystem[C]. Proceedings of the 2022 ACM Conference on Fairness,Accountability,and Transparency,2022.
[25]毕秀玲,陈帅. 科技新时代下的"审计智能+"建设[J]. 审计研究,2019(6):13-21.
[26]龙志能,鲍镕江,何贤杰. 人工智能驱动的审计变革研究[J]. 审计研究,2024(5):53-61.
[27]Dwork C,Minow M. Distrust of artificial intelligence:Sources & responses from computer science & law[J]. Daedalus,2022,151(2):309-321.

[责任编辑:刘　茜]

The Theoretical Connotation, Accountability Boundary, and Framework Construction of Artificial Intelligence Auditing: From the Perspective of Responsible Innovation

CUI Yongmei, YANG Tingyu, YING Wenchi, CAO Runze
(School of Economics and Management, Beijing Jiaotong University, Beijing 100044, China)

Abstract: Artificial Intelligence (AI) has contributed significantly to societal progress, yet it also introduces a range of risks. As an emerging auditing paradigm, AI auditing plays a critical role in promoting responsible innovation in AI. Following the logical progression of "What is AI auditing? What to audit? How to audit?" this study clarifies the theoretical connotation of AI auditing, explores its accountability boundaries, and organizes the auditing subjects, principles, and implementation pathways to construct a comprehensive AI auditing framework. The study identifies three key findings. Firstly, AI auditing refers to "auditing AI" and serves as a mechanism for ensuring accountability in AI's responsible innovation. Secondly, AI auditing focuses on the accountability and auditability of AI, clarifying the risks and accountability boundaries across the AI innovation chain. The auditing scope is centered on AI models, extending inward to data and algorithms and outward to products and ecosystems. Thirdly, AI auditing unfolds through three dimensions: auditing subjects, auditing principles, and implementation pathways. Internal auditing, social auditing, and national auditing each play distinct roles, addressing technical reliability, safety compliance, and ethical objectives across the entire process, from audit planning to execution and completion. By integrating the perspective of responsible AI innovation with the mechanisms of audit accountability, this study constructs a localized and contemporary AI auditing framework, contributing to the development of a uniquely Chinese independent auditing knowledge system.

Key Words: artificial intelligence auditing (AI auditing); artificial intelligence (AI); responsible innovation; auditing framework; auditing accountability