审计理论与方法

从概念到制度:算法审计问责体系研究

张宝山[1],张永忠[2]

(1.东莞理工学院 法律与社会工作学院,广东 东莞 523808;2.华南师范大学 法学院,广东 广州 510006)

[摘　要]随着算法系统深度嵌入社会生活,算法风险与危害日益凸显,建立有效的问责机制成为算法治理的迫切需求。算法审计问责体系旨在对算法技术风险及社会影响进行综合评估,实现对算法主体的有效监督。该体系包含三个核心构件:由审计执行主体、审计客体和报告使用者组成的主体框架;形成国际—国家—行业—算法主体多层级规范的伦理标准体系;构建包含合规责任、可审计性和归责纠偏在内的责任机制。从社会技术系统理论视角出发,算法审计应超越纯技术视角,系统考量技术实现与社会影响的互动关系,形成面向社会公众的问责体系。《个人信息保护合规审计管理办法》为这一体系提供了初步法律依据,但仍需如《人工智能法》等专门立法对算法审计问责制度予以系统规定,确保算法系统的透明、公正和负责任运行。

[关键词]社会技术系统;算法审计;算法问责;伦理标准;责任机制;问责机制;算法治理

[中图分类号]F239;D912.29;DF01　　[文献标志码]A　　[文章编号]2096-3114(2026)02-0034-11

一、引言

算法已深度嵌入社会生活各领域,成为社会运转的基础设施。与此同时,算法歧视、操纵、隐私侵犯和误判等风险也日益凸显,亟须有效的监督约束机制。2012 年哈佛大学教授 Sweeney 通过实证研究发现,谷歌广告算法对黑人群体常用姓名联想的逮捕记录比白人高出 25%,揭示了算法歧视问题[1]。2014 年,Sandvig 等人首次在论文中提出了算法审计(Algorithm Audit)概念,将其定义为推断复杂而不透明算法系统运作的检测方法[2]。目前,算法审计已成为揭露和减轻算法决策系统相关危害的有力武器。联合国《人工智能伦理问题建议书》、美国《纽约市偏见审计法》和欧盟《数字服务法》均对算法审计提出明确要求。在我国,尽管立法并未直接规定算法审计制度,但《中华人民共和国个人信息保护法》(以下简称《个人信息保护法》)第五十四条与第六十四条规定对个人信息处理活动进行合规审计,间接确立了算法审计要求。这一点在 2025 年 2 月国家网信办发布的《个人信息保护合规审计管理办法》(以下简称《个人信息合规审计办法》)及其附件《个人信息保护合规审计指引》(以下简称《审计指引》)中得到印证,《审计指引》第九条规定“个人信息处理者利用自动化决策处理个人信息的,审计时应当重点评价自动化决策的透明度和结果的公平性、公正性”。

众多学者同样提倡利用审计手段对算法进行规制[3-6]。张涛提出,算法审计是预防和减少自动化决策社会风险的有效手段[7]。罗格斯大学法学院的 Goodman 教授等亦强调,算法审计作为关键的问责机制,能够揭露和缓解算法决策系统可能带来的危害[8]。相较于郑石桥教授从经典审计理论出发对算法审计本质和主体的探讨[6],本文更关注算法审计作为一种问责机制在算法治理中的系统性功能。不仅关注“谁来审计”和“审计什么”,还强调“如何追责”和“如何预防与矫正”。但正如英国数据监管合

[收稿日期]2024-12-26

[基金项目]科技部国家重点研发计划(2022YFC3303200);教育部社会科学规划基金项目(23YJAZH208)

[作者简介]张宝山(1995—),男,江西九江人,东莞理工学院法律与社会工作学院讲师,主要研究方向为算法审计、人工智能治理,邮箱:qq525109412@163.com;张永忠(1978—),男,广东饶平人,华南师范大学法学院教授,博士生导师,主要研究方向为算法审计、人工智能治理。

作论坛在关于算法审计的报告中所述,当前各国的算法审计生态系统仍处于萌芽状态[9]。算法审计与个人信息保护合规审计、算法影响评估等诸多概念之间存在混淆,概念不清晰导致算法审计制度化及其适用存在障碍。更为重要的是,作为一种问责机制,算法审计需要形成完整的问责体系,才能有效发挥其监督和治理作用。本文从社会技术系统理论视角重新审视算法审计问责体系,构建了包含主体框架、伦理标准体系和责任机制三个核心构件的算法审计问责体系,为算法审计从概念走向制度提供了理论框架和实践指引。

二、算法审计问责的概念厘清

概念是主体对客体的抽象化描述,其作用在于特定价值之承认、共识与储藏,减轻后来者实现特定价值的思维负担[10]。概念的厘清有利于正本清源,为构建完整的算法审计问责体系奠定理论基础。

(一)算法审计问责制度的概念演进与界定

算法审计经历了从研究方法到法律问责机制的制度化演变过程,起源于20世纪40至50年代美国用于监测社会歧视的审计方法[11]。当时的审计过程通常由研究人员将实验对象的特定特征随机化,然后派遣这些实验对象到实际场景中测试这些特征对结果的影响。这种社会科学审计方法为后来的算法审计奠定了重要基础。进入人工智能时代后,研究人员将这种方法应用于算法领域。随着算法影响的深入,算法审计发展为制度化问责机制,Ada Lovelace研究所将其定位为算法问责的核心机制之一[12]。

算法审计的制度化过程体现了其作为社会技术系统审计的本质特征。传统技术解决主义认为,仅靠技术解决方案就足以解决可能涉及社会、政治、生态、经济和伦理层面的复杂问题[13],忽视或最小化人类、组织和社会价值观及行为的相关性。这种观点难以应对人工智能时代算法与社会紧密交织所产生的复杂问题。算法系统的异化不仅是技术问题,更是社会问题,需要在技术与社会的交互视角下理解和解决。因此,有必要引入社会技术系统理论,系统分析技术子系统和社会子系统的相互影响[14]。基于此,算法审计问责体系应确保技术与社会系统的适配协调,促进自动化决策与社会环境的共同优化。传统财务审计关注的是"账目是否符合准则",其审计对象(财务数据)与审计标准(会计准则)之间存在明确的对应关系,但算法审计面对的是嵌入社会过程的技术系统,其异化可能产生于数据收集、标签定义、模型部署等多个社会技术交互环节。例如,数据收集涉及谁被包含在训练样本中,标签定义涉及选择何种代理指标,模型部署涉及算法输出如何被人类决策者使用。审计不仅要检查算法模型本身,还要审查训练数据的社会代表性、标签中隐含的价值判断,以及算法在组织决策流程中的实际作用。

这种理论转向确保算法审计问责体系能够真实评估算法的社会影响,而非仅仅验证技术合规性。算法审计问责制度将这种审计活动纳入法治框架,通过建立运行体制、伦理标准和责任机制,实现对算法系统的持续审查评估[3]。这一制度体系不仅要明确"由谁监督""监督什么""如何监督"的问题,还要解决"出了问题谁负责""如何追责"等问题,确保算法系统在整个生命周期内的透明、公正和合规。算法审计问责是算法审计制度化发展的结果,问责是算法审计从研究方法走向制度机制的关键环节,它依托算法审计过程与结果进行责任追究,为算法主体的责任认定和承担提供了更加科学的依据①。

(二)算法审计问责制度的关联概念辨析

算法审计与算法型审计工具、算法技术审计、算法影响评估和个人信息保护合规审计等概念紧密相连,易导致概念混淆,不利于算法审计问责体系的构建。下文通过对相关概念进行系统性对比,明确算法审计问责制度的独特属性,避免实践中的认识混淆。

---

①本文认为,算法审计问责是算法审计制度化的必然结果。正如Sandvig最初提出算法审计时仅将其视为检测方法,而Ada Lovelace研究所已将其明确定位为问责机制,这一演变过程表明,问责是算法审计从研究方法走向制度机制的关键转折点。因此,本文所构建的问责体系必然涵盖审计的主体、客体和程序等基本要素,这些要素是实现有效问责的前提和基础。

1. 算法审计问责制度与算法型审计工具

算法型审计工具是传统财务审计中使用的智慧化组件,如大语言模型辅助审计判断。2001 年国务院办公厅《关于利用计算机信息系统开展审计工作有关问题的通知》已提及将高新技术应用于审计工作。与之不同,算法审计问责制度关注对被审计单位所控制算法的审计,二者的主要区别在于算法是作为工具还是审计对象,以及由谁控制。算法型审计工具是提高审计效率的技术手段,而算法审计问责制度是规范和监督算法运行的制度安排,旨在确保算法决策者对其系统的设计、部署和影响负责。

2. 算法审计问责制度与算法技术审计

算法技术审计是狭义的技术审计方法,正如前文所述,算法技术审计是算法审计制度的起源。Raji 和 Buolamwini 将其定义为收集并分析固定算法或模型的输出结果,通过模拟用户群体等发现问题模式的过程[15]。这类审计主要由研究机构开展,不一定通知被审计单位,费用往往由研究机构自行承担。正是从事算法技术审计的研究人员不断发现算法中的异化问题,才使得更多人关注算法治理,呼吁并推动算法审计走向制度化和问责化。然而,算法技术审计与完整的算法审计问责体系存在明显差异。算法技术审计往往忽视人工智能系统运行过程中的治理框架,算法决策的形成涉及人、技术、管理环节等诸多要素,需要更加系统和全面地进行审计,才能揭示其中的问题并予以改正。此外,研究机构进行审计还存在审计成本问题。相比之下,算法审计问责制度整合了技术审计和治理框架审计,通常由法律规定,要求被审计单位接受审计并支付费用。作为制度化的问责机制,算法审计问责体系不仅关注问题发现,还重视责任追究,具有法律强制力,能更有效地推动算法主体采取改进措施。

3. 算法审计问责制度与算法影响评估

算法影响评估是帮助机构了解和降低自动决策系统风险的评估工具,借鉴自环境影响评估制度[16]。作为算法问责的不同路径,算法影响评估与算法审计问责制度存在显著差异。算法审计注重在算法部署后由独立第三方定期开展,而算法影响评估往往是算法主体在部署前自行开展的评估活动。《个人信息保护法》分别规定了个人信息合规审计与个人信息保护影响评估制度,并指出个人信息处理者应当事前进行个人信息保护影响评估。这表明在我国制度语境下,算法影响评估与算法审计制度是截然分开的两种制度工具。从问责效力来看,算法影响评估在算法部署前由算法主体自行开展,是一种较弱的审查形式。例如,谷歌曾提交过一份表明其公司完全符合规定的隐私评估报告,但在此期间,谷歌却多次因违反联邦窃听法而遭到法院裁决[17]。这种自我评估模式容易导致利益冲突,降低评估的客观性和有效性。相比之下,算法审计问责制度往往是指在算法全生命周期内由独立第三方定期开展的审计检查活动,算法审计可能在算法设计时就邀请审计人员开展审计,但主要是指算法运行之后的检查①。其独立性、专业性和持续性特点,使其成为更为有力的问责机制,能够客观评估算法实际运行效果,发现潜在问题,并对算法主体形成有效约束。但两种机制并非对立关系,而是可以互相补充的问责工具。算法影响评估作为事前预防机制,可以帮助算法主体提前识别潜在风险,而算法审计问责体系则作为事中事后的监督机制,确保算法主体持续履行其责任义务。

4. 算法审计问责制度与个人信息保护合规审计

算法审计问责制度与个人信息保护合规审计之间既有区别又存在紧密联系。个人信息保护合规审计侧重个人信息处理活动,而算法审计问责制度不仅关注个人信息处理合规,还关注算法的伦理合规,更关注平等、自由等基本人权。个人信息保护合规审计主要聚焦于个人数据的合法处理,是对数据治理的问责,而算法审计问责体系则着眼于算法决策的全面影响,是对算法治理的综合问责。但两者又存在紧密联系:对于算法审计问责制度而言,由于个人信息合规是审计关注的对象之一,因此个人信息保护

①值得注意的是,算法审计亦存在内部审计,尽管如此,内部审计也存在独立性要求(即内审团队与算法开发团队应当是彼此分开的两个团队),但一定程度上,内部审计团队可以承担算法影响评估的职责,算法影响评估可能是内部审计活动的一部分。

合规审计活动是算法审计的基础。而由于算法使用个人信息亦是个人信息处理活动之一,因此个人信息保护合规审计中同样会涉及算法审计。这种交叉关系意味着两种问责机制可以共享审计方法、标准和经验,形成协同效应。因此,《个人信息合规审计办法》的出台为我国信息科技治理领域合规审计体制机制奠定了基础,其中关于审计主体、程序和责任的相关规定,为构建算法审计问责体系提供了有益参考,对此下文将进一步论述。

从上述概念辨析中(见图1),可总结算法审计问责制度的核心特征:首先,作为社会技术系统审计制度,算法审计问责制度是对算法技术风险及社会影响进行综合评估的制度,用以监督问责算法主体,保障社会公共利益和个人权益。这种问责不仅关注技术合规性,更强调算法主体对决策结果的责任承担。其次,为保障问责效力,算法审计主要由独立第三方开展,算法主体有义务配合审计。这种独立性确保审计结果不受利益干扰,是问责机制有效运行的前提。最后,算法审计是一种综合性审计。因算法和社会环境均动态变化,社会技术系统审计需建立灵活的审计流程,重视社会与技术因素的相互作用,真实评估算法的社会影响。综上所述,算法审计问责制度是一种独立性强、覆盖全面、运行完整的问责机制,它通过对算法系统全生命周期进行监督检查,确保算法主体对算法设计和运行结果承担相应责任,从而实现算法系统的透明、公正和负责任运行。

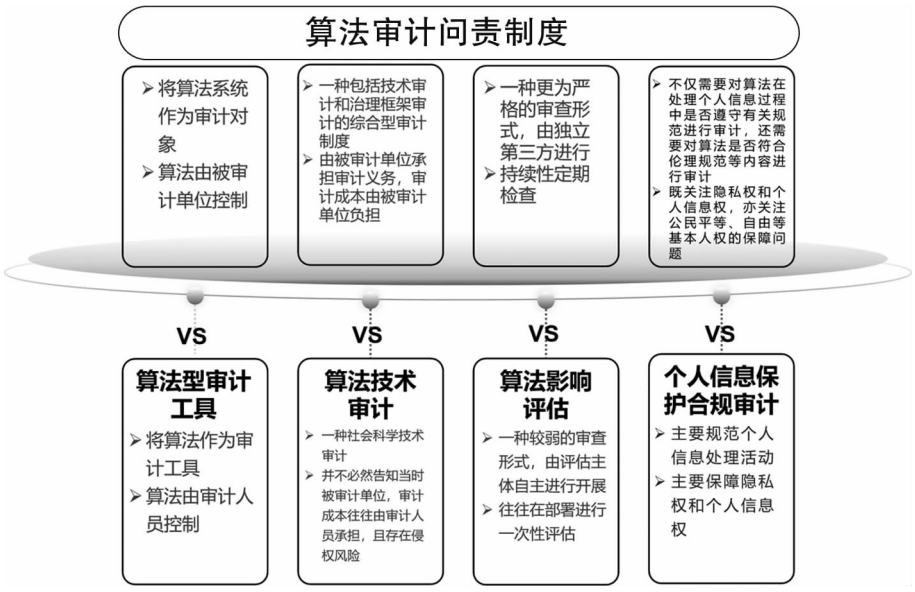

图1　算法审计问责制度关联概念比较

三、算法审计问责的运行体制

算法审计问责体系的构建不仅要明确“问责什么”(概念厘清),还要解决“谁来问责”“向谁问责”的主体框架问题。运行体制是问责得以实施的组织基础:审计执行主体的选择决定了“谁来追责”,审计客体的厘定明确了“向谁追责”,审计报告的使用与披露则为责任追究提供证据支撑并实现透明监督。这些要素共同回应了问责过程中的主体定位和程序保障问题。

(一)审计执行主体的选择

既有实践中,开展算法审计业务(研究)的一般有以下三类机构:(1)会计师事务所。会计师事务所负责传统财务审计业务,并且可能会接受被审计单位委托开展内部控制审计与信息系统审计活动,具备较强的从事审计业务的经验。目前,毕马威(KPMG)和德勤(Deloitte)等大型会计师事务所均开发了算法审计的服务。(2)社会研究机构。包括一些高校、专门从事算法审计研究的公益研究机构以及从事算法合规业务的单位,例如纽约大学设立的 NYU Ad Observatory、Ada Lovelace 研究所、凯西奥尼尔(ORCAA)咨询公司等。(3)政府机构。政府机构主要负责对提供公共服务的算法进行审计,芬兰、德国、荷兰、挪威和英国最高审计机构共同出台的《机器学习算法审计白皮书》,对公共机构的算法审计实践提供指引。2022 年 3 月,荷兰政府公布了一项审计结果,对政府机构使用的九种算法进行了审查,并

公布六种算法未通过审计[18]。除上述三类外，算法主体的内部审计部门亦可以作为审计执行主体，但由于内部审计存在利益冲突问题，内部审计部门应主要作为外部审计的补充。

本文认为，算法审计作为新兴事物，对其采取较为开放的态度更有利于行业的发展。同时，社会技术系统审计方法亦要求更多来自技术、法律和社会科学等跨学科跨领域专家共同参与，以识别和检测算法中复杂的不公正问题。因此，我国可以考虑采取认证或审批制的形式，若具有算法审计的专业资格，无论会计师事务所抑或社会研究机构均应获准开展算法审计业务。《个人信息合规审计办法》规定个人信息合规审计由个人信息处理者内部机构或者委托专业机构进行，并鼓励专业机构按照《中华人民共和国认证认可条例》的有关规定进行认证，表明我国并未对审计执行主体资质限定在某一行业。但注重行业发展的同时，应强调审计机构独立性，建立独立性审查监督机制，避免形式化审计[19]。

（二）审计客体的厘定

审计客体是接受审计的经济责任的承担者和履行者，即被审计单位。由于人工智能价值链的不同阶段可能涉及不同的参与者，因此需要明确审计义务人。判断由谁承担审计义务和责任的关键在于谁控制着算法，控制者有权决定算法的目的和设计，是算法所产生的社会影响的直接责任人，应当承担审计义务，接受监督和问责。《个人信息保护法》未采用《通用数据保护条例》（以下简称 GDPR）的控制者和处理者表述，而统一使用“处理者”，这对个人信息保护并无不妥，但算法审计问责应确立控制者角色。值得注意的是，控制者并非唯一的，如果两个单位均对算法拥有控制权，则构成共同控制者。

控制者的识别重点需关注谁对算法开发目的和设计过程拥有决定权，包括决定：（1）用于训练算法模型的数据的来源和性质；（2）模型的目标输出（预测或分类的内容）；（3）将用于从数据中创建模型的多种机器学习算法类型（如随机森林、神经网络等）；（4）模型需要使用的特征；（5）关键模型的参数；（6）评估指标；（7）模型后续如何进行持续测试或更新；（8）算法模型部署运营后的风险控制安排（如检测、更新等）[20]。这些决策权影响算法行为和结果，构成问责基础。若参与者无权决定这些事项，仅根据合同提供服务，则不承担审计义务，但需配合审计过程。

需要说明的是，并非所有算法控制者都需要接受同等程度的审计。应建立风险导向的审计机制，根据风险等级设计不同审计要求[3]。风险分级是问责体系中资源优化配置的重要机制，能够使问责力度与风险程度相匹配。《个人信息合规审计办法》第四条规定：“处理超过 1000 万人个人信息的个人信息处理者，应当每两年至少开展一次个人信息保护合规审计。”第五条则规定了三类情形下保护部门可要求专业机构进行合规审计：存在较大风险、可能侵害众多个人权益、发生重大安全事件的情况。这种量化指标与风险情形相结合的分级思路，可为算法审计问责体系的风险分级提供参考。在算法审计问责体系中，风险分级可基于以下因素：（1）算法决策对象的规模（如影响人数）；（2）算法决策的敏感性（如是否涉及重大权利、敏感信息等）；（3）算法自主性程度（如人类监督的强度）；（4）算法透明度（如可解释性高低）；（5）算法应用领域（如医疗、金融、司法等关键领域）。风险等级越高，问责要求越严格，审计频率越高①。

（三）审计报告的使用与披露

审计报告是算法审计问责体系的核心成果，其使用和公开直接关系到问责的实际效果。问责的本质在于“向相关利益方说明情况、解释原因并承担后果”，因此审计报告的适当传达和有效使用是完成问责闭环的关键环节。审计报告的用户包括被审计单位、监管机构和社会公众，只有当审计发现的问题能够得到有效传达并促成实际改变，问责闭环才能真正形成。算法审计报告的预期使用者主要包括三类：首先，被审计单位需根据报告建议调整算法模型和内部治理框架，是首要使用者。《个人信息合规

①明确审计客体及其风险分级，是实现精准问责的前提。通过确定“向谁问责”并根据风险程度差异化问责强度，体现了问责的针对性和比例原则。

审计办法》第十一条规定处理者应对发现问题进行整改,并在 15 个工作日内报送整改报告,这种机制是问责效力的重要保障。其次,网信办作为监管机构,可将审计报告作为问责依据,体现问责的惩戒功能。最后,用户和受影响群体作为利益相关方,拥有知情权和算法解释权,让他们了解审计结果是促进社会监督和算法透明的重要途径。

关于审计报告的披露,基于推动决策系统的透明应当体现合比例性原则与价值平衡需求[21],在算法主体可承受的规制负担内披露审计报告较为合适。例如欧盟《数字服务法》第二十八条规定了审计报告的内容,指出需要在报告中说明审计的主要结果,以此降低审计活动对算法商业秘密的泄漏风险。本文建议应当要求个人信息处理者对外披露审计结果,为保护商业秘密,可不披露审计过程,但基于监管的需求,可能需要向政府部门提供完整的审计线索。问责需要透明度,但并非无限度的透明。在算法审计问责体系中,报告披露应当遵循"分级分类"原则:向监管机构提供完整版报告,包含详细的审计过程、发现的所有问题及相关证据;向公众披露摘要版报告,包含主要问题发现、整改措施和合规状态评价;针对特定严重问题,可向直接受影响群体提供相关部分的详细说明。同时,应建立审计报告的标准化格式,确保披露内容的一致性和可比性,增强公众对审计结果的理解和信任。

四、算法审计问责的伦理标准体系

明确"谁来问责""向谁问责"后,还需解决"依据什么标准问责"和"如何预防问题"。构建科学合理的算法伦理标准体系是实现有效问责的关键,在问责语境下,算法伦理标准体系具有双重功能:一方面,它为算法开发者提供了行为指引,使问责具有预期效果;另一方面,它为审计人员提供了评判依据,使问责具有客观标准,体现了其责任追究功能。

传统财务审计过程中审计人员所依据的参照标准,通常是指企业遵循的会计准则。算法审计同样需要依据一定的标准作为审计依据,用以描述法律的要求、用户的期望等。算法运行过程往往按照工具理性行事,以成本和收益、风险和回报、手段和目的的计算为指导,忽视个人权利、社会公平、民主参与等人文价值。因此,技术设计应当立足于社会制度背景和技术文化根基,通过算法系统的设计来实现对人们行为的助推、引导与转化,使算法技术更好地服务于人类和社会福祉,实现"技术道德化"[22]。从问责角度看,这种"技术道德化"正是通过将价值理性嵌入算法设计和运行过程,使技术主体对其社会责任有所回应,而标准体系则将这种责任具体化和可评估化。算法的规范标准即是对算法伦理进行描述的标准(即算法伦理标准),调整算法与社会之间的伦理关系。换言之,算法伦理标准是问责的尺度,算法审计是问责的过程,两者共同形成问责的核心机制。

我国近年来强调加强人工智能伦理审查,如《新一代人工智能伦理规范》强调在算法开发中加强伦理审查,《科技伦理审查办法》明确建立科技伦理委员会审查算法遵守公平、公正等原则的情况。《中华人民共和国国民经济和社会发展第十五个五年规划纲要》指出,要"完善人工智能领域法律法规、政策制度、应用规范、伦理准则"①,明确将伦理准则作为人工智能治理的核心要素。这些政策文件的出台,为算法伦理标准体系的构建提供了政策基础。

伦理标准由谁制定、参照何种标准进行审计,是问题的关键。澳大利亚《2021 年在线安全法》、英国《在线安全法》和欧盟《数字服务法》主要将行业行为准则作为标准。本文认为,伦理标准制度需要进行体系化建构,依次形成国际—国家—行业—算法主体四层级制伦理规范体系(见图 2)。这种多层级的伦理标准体系,体现了问责的层次性和适应性,使问责既有统一性,又有针对性,能够更好地适应不同情境下的问责需求。

首先,伦理标准需贴合特定区域的公众期望,维护当地价值观念[23]。国际人工智能伦理政策文件

①https://www.news.cn/politics/20260313/085af5de5a4b4268aa7d87d90817df2f/c.html。

均有所体现，如欧洲强调遵循《欧盟基本权利宪章》的价值观，美国强调保护公民自由。伦理标准会因地域而异，但需与国际伦理共识协调。其次，由于算法服务于特定行业，需考虑行业突出问题和地区风俗习惯。行业性标准能够使问责更加精准，避免“一刀切”带来的不适应性问题。最后，算法主体应以“上位法”为依据，制定符合自身情况的实施方案，确保可执行性，如以义务或禁止等易转化为计算机语言的表达方式。可见算法伦理标准体系应具备以下特征：一是可操作性，标准应明确具体，便于执行和评估；二是动态适应性，能够随技术和社会发展而更新；三是层次性，能够涵盖从原则到细则的不同层次；四是可及性，保证相关方都能了解和获取标准内容。

图2　算法的伦理标准体系

构建完善的算法伦理标准体系是问责制度有效运行的前提，算法审计将依据这些伦理标准，评估算法主体是否遵循了上位法规定，并检查其伦理标准的实施情况。这一过程不仅促进了算法的透明性、公正性和合规性，而且保障了算法技术能够在尊重个体与群体权益的同时，发挥其在社会中的积极作用。

五、算法审计问责的责任框架

问责体系的最终落脚点在于责任的明确与追究。在确立主体框架和标准依据后，本部分直接回应“如何追责”和“如何矫正”的核心问题。

（一）算法主体的合规责任体系

本文所指责任是广义上的责任，既包括本来意义上的主体职责、义务，也包括事后对违规者或违约者追究的法律后果。审计是获取与审计事务相关的证据，评价其与既定标准的符合程度，得出审计结论的系统过程。换言之，审计检查被审计单位是否按标准从事相关活动，合规是审计活动的前提之一。在算法审计问责体系中，合规责任是最基础的责任形式，强调算法主体事前主动遵守规范标准。这种责任具有预防性特征，通过明确主体义务促使算法主体自觉规范行为，从源头减少风险。

前文已论述算法伦理标准体系，算法主体应据此设置适用于自身的实施方案，包括依据上位法建设隐私保护程序、设置算法参数、调整机器学习误差与偏差及对工作人员的伦理培训等。检查实施方案是审计的第一步。为保证算法主体重视方案建设与执行，有必要将合规责任细化到个人。一般而言，企业高管对于企业合规负有直接责任，例如美国特拉华州最高法院在 Stone v. Ritter 案中就明确了公司合规监督义务属于董事会的善意义务[24]。算法合规责任应纳入合规责任体系，如 GDPR 规定了数据保护官制度，《个人信息保护法》第五十二条则同样规定了一定规模以上的个人信息处理者应当指定个人信息保护负责人①，对个人信息处理活动进行监督。

在我国，首席数据官制度正逐渐在企业和政府机构中得到推广和实践。江苏省通过发布《企业首席数据官制度建设指南》来推动企业建立首席数据官制度，而《广东省数据要素市场化配置改革行动方案》则提出了政府部门首席数据官制度的试点。这表明首席数据官作为一个重要的职位，不仅适用于私营企业，也适用于公共部门，其跨领域的适用性显示出对当前数据驱动需求的积极响应和适应性。我

①尽管个人信息保护负责人与首席数据官、数据保护官等在相关法律条文的表述上有所不同，但从职位的职责设置来看，存在内在一致性，为尽可能还原相关法律条文中的直接表述，本文将对以上几个概念进行交叉使用，特此说明。

国需考虑将算法合规责任纳入首席数据官职责当中,形成两层责任体系:

第一层为组织层责任。董事会应当通过设立首席数据官的方式履行合规义务,公共服务部门同样应当将负责信息安全的主管领导确立为首席数据官,以此保障算法合规性。欧盟 GDPR 第三十七条要求,公共机构以及需要大规模地对数据主体进行常规和系统性监控,或者处理敏感个人数据及有关刑事定罪和犯罪相关个人数据的控制者或处理者,应当指定数据保护官。2019 年德国汉堡数据保护和信息自由部门就针对 Facebook 位于德国的子公司未及时设立数据保护官而对其处以 51000 欧元的罚款[25]。相关主体还应当保障首席数据官能够有效履职,例如在 2022 年 LeistritzAG v. LH 一案中,欧盟法院指出控制者或处理者应当保障数据保护官的职权行使,不得随意解雇或处罚数据保护官[26]。美国参议院和众议院发布的《美国数据隐私和保护法案(草案)》当中,同样规定了符合条件的个人信息处理者应当指定隐私和数据安全官,并且规定了隐私和数据安全官配合进行审计的义务。

第二层为专职人员责任。首席数据官应依据其职责定位,构建算法价值观以推动算法向善,同时应配合审计人员开展算法审计工作,并依据审计结果对算法进行进一步的完善。例如 GDPR 第三十八、三十九条规定了数据保护官的职责和任务,其中就提及配合审计的责任以及协助开展数据保护影响评估的义务。这种合规责任不仅包括算法开发启动阶段的合规建设责任(如在项目设计开发初期建立相关实施方案),还包括算法运营过程中的持续性合规监督责任(如算法运行过程中出现偏差时的纠偏责任)。2021 年比利时数据保护局就依据 GDPR 以数据保护官未充分履职为由对 Proximus SA 公司进行了处罚,同时,该案强调应当保证数据保护官履职的独立性,不得分配与数据保护官职务相冲突的工作[27]。首席数据官制度在外部监督之外建立了组织内部的责任约束,有助于形成多层次的问责结构。

(二) 算法主体配合审计的责任

算法的可审计性是指外部各方可以检查和评估它们的行为、性能和影响[28],有助于识别和纠正系统中的错误、偏差和风险,为算法结果提供证据和解释。可审计性不仅是技术要求,更是责任机制,确保问责的可行性和有效性。尽管目前一些组织或研究机构会针对市场上运行的算法是否存在偏见等问题进行检测,但由于算法主体不存在配合的义务,研究结果不会得到回应。例如纽约大学广告观察站于 2020 年 9 月招募志愿者参与非侵入式用户审计,通过安装扩展程序收集他们在 Facebook 上看到的政治广告数据,但启动不到一年,Facebook 就以其“违反服务条款”为由,禁用了广告观察站的所有平台访问权限[29],使得这一公益研究被迫中断。这一事件揭示了当前算法审计面临的核心障碍:缺乏法律强制力要求算法主体配合审计。早在 2016 年,Sandvig 就提起联邦诉讼,要求政府不得根据《美国计算机欺诈和滥用法》(CFAA)追究算法技术审计研究人员的刑事责任,目前该案已获胜,联邦法官认定,为研究目的获取数据而违反服务条款的行为并不构成刑事犯罪[30]。在 Van Buren 诉美国政府案中(Van Buren v. United States),美国联邦法院更是直接指出,“旨在揭示在线算法是否导致种族、性别或其他歧视的研究不违反《计算机欺诈和滥用法》”[31]。但英国数据监管合作论坛报告指出,算法主体拒绝合作导致审计机构难以获得充分证据,阻碍了成熟审计生态系统的形成[9]。因此,有关部门必须通过制度要求算法主体履行配合审计的责任,以增加算法的可审计性。《个人信息合规审计办法》第八条明确规定个人信息处理者应当为专业机构正常开展个人信息保护合规审计工作提供必要支持,正是对算法主体配合审计责任的制度化体现。从问责角度而言,这种配合义务是一种程序性责任,确保问责过程的顺利进行。

可见,算法的可审计性是确保算法决策过程透明、公正和负责任的关键要素。基于社会技术系统理论框架,算法的可审计性不仅关乎技术层面的透明度,也涉及社会层面的责任和伦理。在对以上内容进行审计时,需要取得被审计单位的训练数据集、被审计单位保存算法设计与训练日志,以及被审计单位提供用户信息以确定算法运行背景及审计的范围等,若未确定被审计单位配合审计的义务,难以取得以上审计证据,则审计工作将难以达到预期的效果。

诸多国家(地区)已在立法中强调了算法系统的可审计性。欧盟《人工智能法》第十二条规定保证系

统可追溯性，要求记录系统表现，并赋予当局查阅权力。美国审计署（GAO）发布的《联邦信息系统控制审计手册》涵盖数据质量、代表性、算法设计等多方面审计，要求被审计单位提供完整信息并保持良好沟通。

总而言之，可审计性可以从技术和伦理两方面评估算法、模型和数据集：技术方面侧重用于衡量和测试算法功能和质量的方法和工具，例如准确性、鲁棒性及安全性等；伦理方面需要提供检测算法设计和使用的价值观和原则的渠道，例如隐私、透明度、责任和人为监督等。为落实针对这两方面的审计，算法审计除了询问、观察、检查、重新执行等传统审计程序外，实践中还出现了针对算法所设计的方法，例如前文提及的代码审计、虚拟用户审计等。目前算法审计创新主要在技术方法层面，有待上升到制度层面，而可审计性是实现这一跃升的基本路径。通过法律制度确立算法主体的配合审计责任，能够打破“黑箱”壁垒，使算法系统的运行和决策过程变得可检查、可评估、可追责，从而实现真正的算法问责。

（三）审计不合格的归责与纠偏

归责和纠偏机制保证了算法问责的约束力和执行力。在社会技术系统理论视角下，审计不合格的归责和纠偏不仅是法律责任认定，也是对社会价值观和伦理标准的回应。审计不合格可能产生的责任包括算法优化责任和算法侵权责任。

算法优化责任指因算法自主学习或数据本身携带偏见因素造成算法异化时，算法主体虽无须承担侵权责任，但有责任纠正这种异化。这种责任体现了问责的改进功能，其重点不在于惩罚，而在于纠正问题、完善系统。当算法出现严重偏差时，审计人员应及时与被审计单位管理层沟通，欧盟《数字服务法》第三十七条规定被审计单位应在收到非肯定性报告后一个月内执行审计建议。如果被审计单位经沟通仍然拒绝改正的，审计人员应当出具“否定意见”的审计报告并向有关机关报告审计结果，由有关机关对相关问题进行处理或问责。《个人信息合规审计办法》第十一条同样规定了个人信息处理者按照保护部门要求对合规审计中发现的问题进行整改的责任。

算法侵权责任则是指算法因不当的设计、训练、运用给个人权益或者公共利益造成损害的，算法主体所须承担的侵权责任。这种责任体现了问责的惩戒功能，其目的在于对已造成的损害进行追责和赔偿，同时通过惩罚震慑类似行为。算法侵权可能会产生行政及民事责任。就行政责任而言，我国诸如《个人信息保护法》（第七章）规定了数据处理者在违反法律关于数据保护义务时的行政责任。欧盟与美国同样为算法设置专门的“行政处罚”，例如欧盟《人工智能法》规定设置禁止类人工智能（第五条）、撤回违法人工智能（第二十四条），以及美国一些州政府发出某些算法技术的禁令（例如加利福尼亚州就出台法规禁用警方随身携带式面部识别技术）[32]。除行政处罚外，算法侵权还可能因对个人权益或公共利益造成损害而触发民事责任。我国《个人信息保护法》第六十九条就明确规定个人信息处理者对个人权益造成损害的应当向个人承担赔偿责任，同时第七十条规定了人民检察院、法律规定的消费者组织和由国家网信部门确定的组织可以在算法侵害众多个人权益时提起公益诉讼。在侵害问责的过程中，审计报告可以作为认定责任的重要证据，以此推动公众参与算法价值观的纠偏。这种以审计为基础的责任认定机制，为算法问责提供了事实依据，增强了问责的客观性和可信度。

综上，审计不合格的归责要求算法主体在技术发展的同时反思其社会影响，确保技术进步与社会价值协调。通过构建以审计为基础、多种责任形式并存、公众广泛参与的追责机制，可有效推动算法主体履行社会责任，促进算法系统健康发展。

六、结语

本文通过“概念厘清→运行体制→伦理标准→责任框架”的逻辑链条，系统回应了算法审计问责的核心问题：概念厘清明确了审计问责制研究的对象和范围；运行体制确立了“谁来问责”“向谁问责”的主体框架；伦理标准为“依据什么问责”和“如何预防”提供了规范依据；责任框架则直接落实了“如何追责”和“如何矫正”的具体机制。随着人工智能技术的不断进步，算法审计问责体系将在确保算法公正、透明和负

责任方面发挥越来越重要的作用。《个人信息合规审计办法》对于审计主体、客体以及程序等规定,对我国信息科技领域审计制度体系建设起到了奠基作用。这一法规的出台标志着我国信息科技审计实践进入新阶段,为算法审计问责体系构建提供了重要参考。然而,作为一种专门针对算法系统的问责机制,算法审计问责体系仍需进一步完善。面对人工智能技术的快速发展和应用拓展,现有法律框架难以回应算法伦理的全面需求。因此,亟须制定如《人工智能法》等专门性立法,对算法审计问责体系的主体框架、标准体系以及责任机制进行系统化的规定,构建更加完善的算法问责闭环。同时,如何规范指引算法审计从业人员开展算法审计活动,还有待进一步研究设计相关审计活动标准。从前文的分析当中不难看出,算法审计与传统财务审计分属于审计规范体系中不同的子系统,还应进一步研究和设计适应算法审计特点的活动标准。但我国出台的一些用于财务审计的规范,例如《第 2203 号内部审计具体准则——信息系统审计》等,能够对算法审计工作开展和制度设计起到指导作用,应当作为算法审计规范制定过程中的参考依据。本文主要关注算法审计问责制度体系的整体构建,但在此体系下的诸多细节,如伦理标准体系各层级间的协调机制以及如何精细化构建问责过程中的证据链和责任追究程序等方面,尚有待后续研究进一步深入探讨。

参考文献:

[1]Sweeney L. Discrimination in online ad delivery[J]. Communications of the ACM,2013,56(5):44-54.

[2]Sandvig C, Hamilton K, Karahailos K, et al. Auditing algorithms: Research methods for detecting discrimination on internet platforms[C]//Proceedings of the Data and Discrimination Workshop,2014:4349-4357.

[3]张永忠,张宝山. 算法规制的路径创新:论我国算法审计制度的构建[J]. 电子政务,2022(10):48-61.

[4]Lam K, Lange B, Blili-Hamelin B, et al. A framework for assurance audits of algorithmic systems[C]//Proceedings of the 2024 ACM Conference on Fairness, Accountability,and Transparency. New York:ACM,2024:122-130.

[5]贺勇,李佳蔚,刘筱祎. 数字平台算法审计:现实理据、客观挑战与关键进路[J]. 南京审计大学学报,2025(2):10-20.

[6]郑石桥. 算法审计本质论[J]. 财会月刊,2024(15):73-77.

[7]张涛. 通过算法审计规制自动化决策以社会技术系统理论为视角[J]. 中外法学,2024(1):261-279.

[8]Goodman E P, Tréhu J. AI audit-washing and accountability[EB/OL]. (2022-11-15)[2025-06-07]. https://www.gmfus.org/publications/ai-audit-washing-and-accountability.

[9]GOV.UK. Auditing algorithms, the existing landscape, role of regulators and future outlook[EB/OL]. (2022-09-23)[2025-06-07]. https://www.gov.uk/government/publications/findings-from-the-drcf-algorithmic-processing-workstream-spring-2022/auditing-algorithms-the-existing-landscape-role-of-regulators-and-future-outlook .

[10]黄茂荣. 法学方法与现代民法[M]. 北京:中国政法大学出版社,2000:52.

[11]Gaddis S M. Audit studies: Behind the scenes with theory, method,and nuance[M]. Cham: Springer,2018: 7-17.

[12]Ada Lovelace Institute, AI Now Institute, Open Government Partnership. Algorithmic accountability for the public sector[EB/OL]. (2022-06-20)[2025-06-07]. https://www.opengovpartnership.org/documents/algorithmic-accountability-public-sector/.

[13]Morozov E. To save everything, click here: The folly of technological solutionism[M]. University Park: Penn State University Press, 2013: 5-6.

[14]Clavell G G, Aumaitre A, Calders T. How a socio-technical approach to AI auditing can change how we understand and measure fairness in machine learning systems[C]//AI-based multimodal medical examination system. Cham: Springer,2024:125-135.

[15]Raji I D, Buolamwini J. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products[C]. Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, 2019: 429-435.

[16]Kaminski M E, Malgieri G. Algorithmic impact assessments under the GDPR: Producing multi-layered explanations[J]. International Data Privacy Law,2021,11(2):125-144.

[17]Hoofnagle C J. Assessing the federal trade commission's privacy assessments[J]. IEEE Security & Privacy,2016,14(3):62-71.

[18]Netherlands Court of Audit. An audit of 9 algorithms used by the Dutch government[EB/OL]. (2022-05-18)[2025-06-07]. https://www.courtofaudit.nl/en/publications/an-audit-of-9-algorithms-used-by-the-dutch-government.

[19]Engler A C. Independent auditors are struggling to hold AI companies accountable[EB/OL]. (2024-03-02)[2025-06-07].

https://www.fastcompany.com/technology/independent-auditors-are-struggling-to-hold-ai-companies-accountable.
[20] ICO. Guidance on the AI auditing framework[EB/OL]. (2022-02-14)[2025-06-07]. https://ico.org.uk/media/2617219/guidance-on-the-ai-auditing-framework-draft-for-consultation.pdf.
[21] 张永忠. 论人工智能透明度原则的法治化实现[J]. 政法论丛,2024(2):124-137.
[22] 朱羿锟,张宝山. 社会信用治理模型选择:由风险预测制迈向行为积分制[J]. 求实,2024(4):68-85.
[23] 张永忠. 算法价值观的法治化建设研究[J]. 法治研究,2025(6):76-88.
[24] Bainbridge S M, Lopez S, Oklan B. The convergence of good faith and oversight[J]. UCLA Law Review,2008,55(3):559-604.
[25] Hamburg Data Protection Commissioner. 51,000 fine against Facebook Germany gmbh[EB/OL]. (2019-10-21)[2025-06-07]. https://edpb.europa.eu/news/national-news/2019/hamburg-data-protection-commissioners-eu51000-fine-against-facebook-germany_en.
[26] Court of Justice of the European Union. Case C-534/20(Leistritz)[EB/OL]. (2022-06-22)[2025-06-07]. https://www.dpcuria.eu/case?reference=C-534/20.
[27] Gegevensbeschermingsautoriteit. Dossiernummer: AH-2019-0013[EB/OL]. (2020-04-28)[2025-06-07]. https://www.gegevensbeschermingsautoriteit.be/publications/beslissing-ten-gronde-nr.-18-2020.pdf.
[28] Diakopoulos N, Friedier S, Arenas M, et al. Principles for accountable algorithms and a social impact statement for algorithms[EB/OL]. (2018)[2025-06-07]. https://www.fatml.org/resources/principles-for-accountable-algorithms.
[29] Bobrowsky M. Facebook disables access for NYU research into political-ad targeting[N]. The Wall Street Journal,2021-08-04(B3).
[30] Urman A, Smirnov I, Lasser J. The right to audit and power asymmetries in algorithm auditing[J]. EPJ Data Science,2024,13(1):19.
[31] Bhandarie E, Snowden R. ACLU challenges computer crimes law thwarting research on discrimination[EB/OL]. (2020-03-27)[2026-02-26]. https://www.aclu.org/news/racial-justice/aclu-challenges-computer-crimes-law-thwarting-research.
[32] California Legislative Information. Bill text-AB-1215 law enforcement:facial recognition and other biometric surveillance[EB/OL]. (2019-10-09)[2025-06-07]. https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=201920200AB1215.

[责任编辑:高　婷]

From Concept to Institution: Research on Algorithmic Audit Accountability System

ZHANG Baoshan[1], ZHANG Yongzhong[2]

(1. Law and Social Work School, Dongguan University of Technology, Dongguan 523808;
2. School of Law, South China Normal University, Guangzhou 510006)

Abstract: As algorithmic systems become deeply embedded in social life, algorithmic risks and harms have become increasingly prominent, making the establishment of effective accountability mechanisms an urgent requirement for algorithmic governance. The algorithmic audit accountability system aims to comprehensively evaluate algorithmic technical risks and social impacts, enabling effective supervision of algorithm operators. This system comprises three core components: a subject framework consisting of audit executors, auditees, and report users; an ethical standard system forming multi-level regulations across international, national, industrial, and algorithmic operator levels; and a responsibility mechanism encompassing compliance obligations, auditability requirements, and liability determination with corrective mechanisms. From the perspective of socio-technical systems theory, algorithmic audits should transcend purely technical perspectives, examining the interactions between technical implementation and social impacts, forming an accountability system oriented toward the general public. The Administrative Measures for Personal Information Protection Compliance Audits provide preliminary legal basis for this system, but specialized legislation such as a dedicated Artificial Intelligence Law is still needed to systematically regulate algorithmic audit accountability, ensuring transparent, fair, and responsible operation of algorithmic systems.

Key Words: socio-technical system; algorithmic audit; algorithmic accountability; ethical standard; responsibility mechanism; accountability mechanism; algorithmic governance