构建多维度AI审计框架思考

◎ 文/ 林慧涓 陈宋生

一、人工智能对财会领域的影响：机会和威胁

近年来深度学习技术的崛起、神经网络架构的改进、计算能力的增强以及互联网数据的可用性，共同推动人工智能领域的技术革新。这些人工智能技术在财务和会计方面已经展示出巨大的潜力和价值。

首先，在数据分析方面，AI和机器学习算法能够处理大量复杂的财务数据，提供更及时的财务报告。这不仅提高财务报告的质量，也使得财务报告使用者的决策更快速

其次，在风险评估和管理方面，AI技术能够分析历史数据和市场趋势，预测潜在的财务风险，从而助力企业更有效地进行资本配置和风险规避。例如AI算法能够识别信贷风险、市场风险和操作风险等多种类型的财务风险，为企业提供更全面的风险评估。

最后，在自动化和效率提升方面，AI技术也发挥重要作用。通过自动化财务流程，如发票处理、支付和收款等，AI不仅降低人力成本，还提高财务操作的准确性和效率。此外，AI还能够自动执行复杂的财务模型和算法，从而在预算编制、资金分配和投资决策等方面提供更高水平的自动化支持。在合规和审计方面，AI技术能够自动检测财务报告和交易记录中的不规范和异常行为，从而提高财务合规和审计的效率。不仅在财经方面有着广泛的应用前景，而且在医疗诊断、代码生成、语言翻译等多个领域体现其优越的性能。特别是像ChatGPT这样的大型语言模型，在某些基准测试上已经接近人类的性能水平。

然而，AI技术的广泛应用也带来伦理和社会的挑战。Weidinger et al.指出，大型语言模型如ChatGPT存在加剧社会偏见的风险。这些模型在输出内容中可能无意中强化对特定社会群体的偏见性描述，从而加剧现有社会的刻板印象。此外，这些模型还可能被用于恶意目的，包括生成高度复杂和个性化的欺诈行为或进行大规模的欺诈活动。同时，AI的架构存在数据泄露和敏感信息推断的潜在风险，这严重威胁个人隐私。因此，从伦理角度对AI严格分析不仅需要扩展现有的治理框架，还需要在风险评估方法、性能基准和伦理框架方面取得革命性的进展。正是这些挑战需要一种跨学科的方法，将数据科学、管理学、伦理学等多个学科联合起来，以全面审视和解决这些挑战。特别是在人工智能审计方面，需要构建一套全面而有效的审计机制，以确保AI技术的健康、可持续发展。

二、人工智能审计的必要性

随着人工智能(AI)技术在各个领域的广泛应用，其潜在的风险和挑战日益凸显。这些风险不只局限于技术层面，还涉及社会、伦理、法律和经济等多方面。因此，对AI进行全面和深入的审计成为迫切需求，有必要通过分析未经审计的AI模型可能带来主要问题，探讨开展人工智能审计的必要性。AI模型的主要问题：

(一)数据偏见

大型语言模型通常基于大量的文本数据进行训练。如果这些数据包含某种形式的偏见(如性别偏见、种族偏见等)，模型可能会学习并在输出中反映这些偏见。这不仅可能误导用户，还可能加剧社会不平等和歧视，从而对社会和伦理方面产生不良影响。

(二)不准确的信息传播

由于AI模型是基于其训练数据进行预测和决策的，如果训练数据包含错误或过时的信息，模型的输出也可能是不准确的。这在医疗、法律或金融等关键领域尤为危险，因为不准确的信息可能导致严重的后果，包括误诊、误判或财务损失。

(三)安全性问题

未经审计的AI模型可能更容易受到对抗性攻击。这些攻击旨在

通过输入特定的数据来误导模型，从而产生不准确或出现误导性的输出。这不仅威胁到模型的可靠性和有效性，而且可能导致更广泛的安全风险。

(四)法律和伦理风险

AI模型在生成内容时可能触犯法律和伦理规定，如侵犯版权或涉嫌诽谤。使用这些未经审计的模型的个人或组织可能因此面临法律责任。此外，模型可能生成不道德或不合法的内容，从而引发更广泛的伦理和社会问题。

(五)可解释性和透明度

由于未经审计的AI模型通常缺乏可解释性，用户难以理解模型的决策逻辑和依据。这不仅影响用户对模型的信任，还可能导致错误或不合理的决策。

(六)社会和文化影响

AI模型可能会生成与特定文化或社会观念不一致的内容，从而引发文化冲突或误解。这不仅可能导致模型在某些文化或社会环境中的应用受限，还可能加剧社会分裂和冲突。

因此，审计在AI模型的治理体系中被认为是一个重要且务实的工具。这一观点不仅得到学界的重视，而且受到高科技企业和政策制定者的重视。例如，欧洲委员会正在考虑将大型语言模型归类为“高风险AI系统”，这意味着设计这些模型的技术供应商必须接受“独立第三方参与的符合性评估”，即另一种名为审计的活动。

AI审计的早期研究主要集中在探索和缓解算法应用中可能产生的歧视及其他不良影响，特别是在决策支持系统以及对个体和组织的影响方面。近年来，研究领域逐渐引入“基于伦理的自动决策系统审计”(EBA)这一概念。EBA被构想成一种结构化流程，用于评估某实体(无论是现存还是历史)的行为是否与相关原则或规范相符。与此相对，Brown et al.(2021)提出的伦理算法审计更侧重于评估算法对利益相关者产生的实际影响，并据此识别特定算法的情境或属性。两者主要区别在于伦理算法审计更侧重于实际影响，而EBA更注重原则和规范的一致性。此外，伦理算法审计(Brown et al.，2021)明确将算法作为审计的主要对象，而没有保留被审计实体的开放性。

以大型语言模型审计为例，AI审计可细分为模型审计、治理审计和应用审计三个关键阶段，以全面而系统地评估与人工智能相关的各方面风险。

首先，模型审计专注于预训练但尚未发布的大型语言模型技术属性，以揭示可能反映在历史不公正的训练数据集中偏见或其他歧视性做法。该审计应由所有设计和传播大型语言模型的技术供应商强制执行，并产生一份详细报告以概述模型的属性和局限性。

其次，治理审计主要针对技术供应商在设计和传播语言模型过程中的管理和决策机制。该审计阶段由独立的第三方或内部机构执行，旨在核实技术供应商自我报告的准确性，并通常只能访问有限的内部流程.其主要目标是评估语言模型的设计和传播流程，尽管它无法预见所有随着AI系统与复杂环境随时间互动而出现的风险。

最后，应用审计则是针对基于语言模型的具体应用，包括两个子阶段：功能性审计和影响审计。功能性审计评估应用是否本身合法和符合伦理，并与预期用途一致；而影响审计则更关注应用输出如何影响不同的用户群体和自然环境。这三种审计类型各自审计与语言模型相关的不同问题，并在实践中相互关联，构成一个多层次的审计框架，以识别和管理AI在各种应用场景中可能带来的风险。

在AI审计的概念框架和评估方法中，应综合考虑基于原则和影响的两种方法，而不是过早地界定该领域。以当前发展迅速的大型语言模型为例，从治理角度看，大型语言模型带来方法论和规范方面的挑战。它们通常分两个阶段开发与应用。首先是在大型、非结构化文本语料库上进行预训练，其次在较小、特定任务的数据集上进行微调。这使得在没有特定情景下很难评估大型语言模型。此外，它们在特定任务上或在规模上的表现可能是不可预测的。因此，治理和技术审计都有局限性。治理审计不能预见所有随着AI系统与复杂环境随时间互动而出现的风险。并且，历史上专注于在明确定义环境中的特定功能，没有能力捕捉大型语言模型等AI模型在许多下游应用中带来的社会和伦理风险的全貌。因此，尽管设计针对AI模型的审计程序面临诸多实践性和概念性的困难，但这

些困难不应成为不开展AI审计的借口。相反,这些问题应视为警示,即不能期望单一的审计机制能全面捕捉与AI模型相关的所有伦理风险。作为一种治理机制,审计能助力技术提供商识别并可能预防风险,影响AI模型的持续(重新)设计,并丰富有关技术政策的公共讨论。尽管面临多重挑战,审计在确保AI模型的准确性、可靠性和伦理性方面仍具有不可或缺的核心作用。

三、人工智能审计的框架构思

在当今高度数字化的商业环境中,如何避免使用人工智能(AI)带来的风险是一个相对较新且日益重要的研究领域。为此有必要从AI审计的定义、实践应用,以及在全球范围内的法律和政策方面构建人工智能审计框架。

(一)人工智能审计的定义与范畴

审计是一种治理机制,它可以用于监控行为和绩效,并且在诸如财务会计和工人安全等领域促进程序规则性和透明度的悠久历史。因此,就像财务交易可以针对真实性、准确性、完整性以及合法性进行审计一样,AI审计中,AI系统的设计和使用也可以针对技术安全性、法律合规性或与预定义的伦理原则一致性进行审计。AI审计是一种系统性和独立的活动,它涉及获取和评估与特定实体(无论是AI系统、组织、流程,或是这些元素的任何组合)的行为或属性相关的证据,并将评估结果传达给利益相关方。它强调审计活动的系统性和独立性,突出在高度依赖数据和算法的现代商业环境中,审计在治理机制中的重要性。被审实体不局限于算法,而是开放的,可以是一个AI系统、一个组织、一个过程,或者是它们的任何组合.根据其侧重点,AI审计可分为狭义和广义两种。狭义的人工智能审计是以影响为导向,重点在于探测和评估AI系统对不同输入数据的输出。广义的人工智能审计则是以过程为导向,侧重于评估整个软件开发流程和质量管理流程。

(二)实践应用层面:AI审计的应用场景

纽约市的AI审计法(NYC Local Law 144)要求使用AI进行就业决策的公司必须接受独立审计。此外,全球范围内的专业服务公司,如普华永道(PwC)、德勤(Deloitte)、毕马威(KPMG)和安永(EY),也提供AI审计服务,以帮助企业评估和管理与AI应用相关的风险。在法律与政策背景方面,在全球范围内,AI审计逐渐受到法律和政策的关注。尽管全球范围内人工智能(AI)审计的实践与学术研究呈逐渐上升的趋势,但该领域仍然是一个复杂且至关重要的研究领域。它不仅跨越多个学科领域和研究方法论,而且受到各类法律和政策环境的深刻影响。在我国,该领域的研究处于关键阶段。

(三)AI审计框架

随着关于如何审计AI项目的指导方针日益普及,多个国际组织和政府已发布一系列有助于内部审计功能的AI审计框架。

COBIT框架。最新版本的COBIT框架,即COBIT2019,由信息系统审计与控制协会(ISACA)在2018年发布,以取代其前身COBIT5。作为一个“综合性”框架,COBIT在企业信息和技术的治理与管理方面得到了国际认可。该框架包括几乎所有IT领域的流程描述、期望结果、基础实践和工作产品,因此非常适合作为审计AI启用项目时内部审计功能的初始起点。然而,COBIT2019的综合性也是其弱点之一。由于覆盖范围广泛,该框架可能缺乏针对特定AI应用场景的深入指导。此外,其复杂性可能导致实施难度增加。

COSO ERM框架。由赞助组织委员会(COSO)在2017年更新的COSO ERM框架包括五个组成部分和20个原则,为内部审计提供一种集成和全面的风险管理方法。COSO ERM的风险管理方法可以为AI的治理提供指导,并有效地管理其相关风险,以造福组织。然而,COSO ERM主要侧重于风险管理,可能忽视AI治理中的其他关键方面,如伦理和社会责任。

美国审计署(GAO)AI框架。GAO开发并于2021年6月发布的人工智能问责框架旨在“帮助管理者确保AI在政府程序和流程中的负责任使用”。尽管这一AI审计框架主要关注政府使用AI的受托责任,但由于它基于现有的控制和政府审计标准,主要适用于政府组织,可能不适用于私营企业或非营利组织。此外,它可能缺乏对特定AI技术或应用的深入分析。

国际内部审计师协会(IIA)人工智能审计框架。由IIA发布的人工智能审计框架包括三个主要组成部分和七个元素,有助于内部审计在短期、中期或长期内评估、理解和

传达人工智能对组织创造价值能力的影响。该框架主要侧重于审计，可能不足以全面地解决AI治理的复杂性和多维性。

新加坡个人数据保护委员会(PDPC)模型AI治理框架。由新加坡个人数据保护委员会(PDPC)与世界经济论坛(WEF)第四次工业革命中心(4IR)共同创建的模型人工智能治理框架第二版，关注四个广泛领域，作为组织推出AI计划时的基础考虑和措施。该框架主要适用于新加坡的法律和文化环境，可能需要进行本地化调整才能在其他地区有效。这些框架的制定不仅有助于提高审计的准确性和可靠性，还有助于确保其在伦理和社会层面上的责任与可持续性。

四、人工智能审计的建设方向和建议

在AI审计的复杂体系中，内部审计与外部审计各自扮演着不可或缺的角色，共同构建一个多维度的审计框架。内部审计作为企业自我评估和监控的重要机制，其核心价值不仅局限于模型和算法的性能与安全性，而且延伸至对企业内部流程和决策机制的深度审查。从信息透明度和准确性的角度，内部审计有助于确保模型设计和性能的自我评估是准确和可靠的，这一点在企业商业决策和战略规划中具有至关重要的作用。通过内部审计，企业能更精准地了解到模型的优缺点，从而做出更明智和合理的商业决策。此外，内部审计员由于能够访问企业的内部流程，因此能全面评估模型从设计到应用的全过程，这不仅有助于识别和管理潜在的技术和安全风险，还确保模型的设计和实施与企业的整体战略和目标是一致的。

然而，内部审计也面临一系列挑战和局限性。内部审计需要高度重视企业的商业利益和伦理风险，以避免与被审计对象之间存在潜在的利益冲突。例如，如果一个企业的AI模型是由高级管理层直接推动或批准的，内部审计员可能会面临来自上级的压力，隐瞒或忽略模型存在的问题。特别是在面临激烈市场竞争和快速技术发展的环境中，任何对模型性能的负面评价都可能影响企业的市场地位和竞争力。此时，内部审计很难独立发表审计意见，特别是当审计结果与企业的商业目标不一致时。因此，审计中缺乏第三方的问责机制可能会导致决策者忽视或淡化那些可能威胁到商业利益的审计改进建议，这可能影响模型的性能和可靠性，还可能引发一系列伦理和法律问题。例如，如果内部审计报告指出，AI模型存在数据偏见，但是要解决这一问题，将降低模型性能，那么管理层可能忽视或淡化这类问题，以维护企业利益。如果存在性别或种族偏见的信贷评估模型用于实践中，可能触发社会不满，还可能导致企业面临法律风险。

与内部审计相辅相成，外部审计作为第三方独立评估机制，其核心价值体现在全方位、多维度的审查能力。在模型性能评估方面，外部审计不仅对模型在特定任务和数据集上的准确性进行深入评估，而且对模型在不同文化和背景上的适用性进行评价。这一点明确了审计师需要具备跨学科的知识体系和视野。从安全性维度出发，外部审计致力于深度识别和评估模型可能存在的安全漏洞，这包括对抗性攻击、数据泄露和未经授权的访问等问题。这体现了审计工作的严谨性和全面性。在伦理和合规性方面，通过先进的数据分析和算法检测技术，外部审计可识别模型是否存在性别、种族、文化或其他形式的偏见和歧视，并据此提出针对性的改进建议。这一环节不仅体现审计工作的社会责任感，也是对模型公平性和合规性的有力保证。除了技术和伦理两个主要维度外，模型的可解释性和透明度也是外部审计的重要组成部分。这涵盖了模型决策逻辑的可解释性、模型训练和应用过程的透明度，以及与模型相关的所有文档和元数据的完整性和准确性，有助于提升模型的社会可接受度和信任度。从更为宏观的社会和文化影响角度来看，外部审计还需关注模型可能对社会结构、公众舆论及信息传播等方面产生的长期和短期影响。这不仅是对模型影响力的全面评估，也是对其可持续发展性和社会责任的一种体现。

因此，内部审计与外部审计应当相互补充，与其他治理机制协同作用，以实现更全面和有效的AI模型审计。这不仅要求审计员具备跨学科的知识和视野，还要求其能全面评估和管理模型在技术、商业、伦理和社会等多个方面的风险和影响。这一多维度的审计框架为AI模型的全周期治理提供了一个全面

而深入的研究基础。

五、人工智能审计可能带来的影响

AI审计作为综合性的治理机制，其核心目标不会仅局限于提升模型在技术方面的准确性与可靠性，而是更广泛地涵盖社会、伦理、法律及经济等多个维度的标准与期望。从技术角度出发，AI审计通过精密的数据分析和模型验证，能有效地识别并修正模型的潜在缺陷与误差，从而增强模型在具体应用场景中的准确性和可靠性，促使其整体性和稳健性的提升。在社会与伦理层面，AI审计具有减缓数据偏见和歧视现象的功能。通过对模型的数据处理和决策机制进行严格审查，审计活动能揭示潜在的偏见或歧视，并据此提出针对性的纠正措施，对缓解模型在实际应用中可能引发的社会不平等和不公具有至关重要的意义。从透明度和可解释性方面考虑，AI审计通常会产出翔实的分析报告，以明确模型的运行逻辑和决策依据，这不仅增加模型的透明度和可解释性，也有助于提升用户及其他利益相关方对模型的信任度。在法律和合规性方面，AI审计发挥着风险规避和合规保障的作用。通过审计，能确保模型和算法的运行符合相关法律和规定，如数据保护法和版权法，从而降低AI应用可能导致的法律风险，并为构建更完善的合规体系提供坚实的基础。从伦理和社会责任视角看，AI审计确保模型在设计和应用阶段严格遵循公平、透明和责任等伦理原则，这不仅提升模型的社会接受度，还能促进AI的开发和应用实践。

在经济层面，经过审计的AI模型通常表现出更高的运行效率和准确度，从而有助于降低运营成本和提高经济效益。此外，通过减少潜在的法律风险和提升用户及其他利益相关方的信任，模型的经济价值得以进一步放大。从多元利益相关者的视角来看，AI审计在企业、政府、消费者和学术界等多个层面都具有不可忽视的价值.对企业而言，AI审计不仅是品牌信誉建设的有效手段，还是一种战略性资产，能够提供更高的投资同报率，吸引更多资本注入，进而推动企业的技术创新和市场拓展。对政府和监管机构来说，AI审计生成的数据和分析报告为政策制定提供科学依据，有助于更准确地把握AI技术的社会影响，从而制定更合理和有效的法规与政策。此外，AI审计在识别和纠正模型偏见及歧视方面具有重要作用，这不仅促进社会公平和稳定，也是构建和谐社会的必要条件。从消费者和公众角度，AI审计增强模型的透明度和可解释性，有助于保障消费者的知情权和选择权，同时也增强公众对AI应用的整体信任度。在学术界和研究人员层面，AI审计作为一种研究工具，可以深入探究AI模型在社会、伦理和法律等方面的复杂影响，这不仅推动了AI伦理和社会影响等领域的学术研究，还可能促进不同学科和领域之间的交叉合作与知识整合。

六、结论与展望

AI审计的实施，有助于确保人工智能模型、算法及其应用在技术、伦理和社会应用层面达到社会的期望。通过分析数据和检验模型，AI审计能有效提升模型在实际应用中的准确性和可靠性，进而优化运行效率和降低运营成本。然而，面对人工智能，尤其是强人工智能(AGI)的快速发展，审计领域将遭遇更为复杂和多维的挑战。从审计方法论角度看，传统审计手段可能不完全适用于AI审计，因为这些方法通常针对更为简单和专门化的系统。因此，学术界和实践界需要多学科合作，共同研发适用于AI的新型审计方法和工具。总体而言，AI审计将面临多层次、多维度和跨学科的技术与伦理等多方面的挑战，这需要全球范围内的企业、高校与研究机构的通力合作和技术创新，方能保证AI模型的安全性、可靠性和伦理合规性。综合考虑技术、法律、伦理和社会因素，AI审计不仅是技术进步的必要条件，还是实现AI有效应用的关键途径，有助于构建一个更为安全、公平和可持续的AI生态系统。

基金项目：北京理工大学珠海学院实验教学示范中心“智能财经人才实验教学示范中心”(2023006ZLGC)；教育教学改革一重点项目“基于知识图谱的智能会计核心课程群改革计划”(2023008ZLGC)；北京理工大学珠海学院校级课题“预算绩效管理对粤港澳大湾区企业高质量发展的影响研究”(XK-2023-013)；线下一流课程“管理会计模拟”(2023012YLKC)。

作者简介：林慧涓，美国注册管理会计师，北京理工大学珠海学院财务管理系主任、副教授；陈宋生，北京理工大学管理与经济学院教授、博士生导师，全国会计领军人才，北京市优秀人才。

原载《会计之友》(太原)，2023.23.32～37