算法规制的路径创新：论我国算法审计制度的构建

张永忠　张宝山

从智能投顾到电子商务，从网络经济到智能制造，从商业到政府自动化决策，算法的运用越来越广泛和深入。算法作为智能化时代的关键技术，在提升经营管理效率，为社会赋能的同时，也可能产生算法歧视等异化问题，有损社会公平正义，并成为人工智能时代的新型风险源。2016年，Beauty AI公司举行了一次由机器人作为裁判的选美比赛，来自100多个国家的6000多人参加，评判结果获胜者几乎都是白人，少数为亚洲人，没有一位黑色人种。

如何保障算法运行结果的公正，成为国际社会和世界各国重点关注的问题。欧盟委员会制定的《通用数据保护条例》(GDPR)第35条规定数据保护影响评估制度，列明自动化处理(算法)过程必须对数据主体的权利与自由带来的风险进行评估，以及《人工智能法案》(Artificial Intelligence Act)第43、47、48条进一步细化的数据评估规则，规定合规评估制度。美国提起的《算法问责法案》(Algorithmic Accountability Act of 2019)是专门针对自动决策影响评估和数据保护影响评估而出台的，其指出自动决策系统影响评估是指对自动化决策系统和自动决策系统开发过程(包括自动化决策系统的设计和培训数据)进行评估的研究，以评估对准确性、公平性、偏见、歧视、隐私和安全的影响；美国《算法公平法案》(Algorithmic Fairness Act)对算法审计体制予以了规定，并强调增加可审计性。加拿大公布的《自动化决策指令》(Directive on Automated Decision-Making)对算法提出一系列的规制要求，其中都提及了算法审计的问题。在欧盟，由芬兰、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》(Auditing machine learning algorithms：A white paper for public auditors)(以下简称《算法审计白皮书》)，为GDPR下的算法审计活动提供指引。联合国教科文组织在2021年公布的《人工智能伦理问题建议书》第59条中，明确提出会员国应考虑设立有关体制机制负责人工智能伦理影响评估、审计和持续监测工作，确保对于人工智能系统的伦理指导。

我国出台的《个人信息保护法》第五十四条规定："个人信息处理者应当定期对其处理个人信息遵守法律、行政法规的情况进行合规审计"，第六十四条也提及"委托专业机构对其个人信息处理活动进行合规审计"。根据个人信息保护法对于自动化决策(自动化决策是算法技术应用的一种方式，本文所指自动化决策即为算法，另外，算法亦是人工智能决策处理中枢，因此文中人工智能与算法亦属同等概念)的定义："自动化决策，是指通过计算机程序自动分析、评估个人的行为习惯、兴趣爱好或者经济、健康、信用状况等，并进行决策的活动。(第七十三条)"算法对于信息的处理属于个人信息处理者信息处理活动之一。尽管我国《个人信息保护法》对个人信息的定义采用了"保护记录"模式，即强调信息构成记录但不区分是否被自动化处理，但从个人信息保护制度发展的沿革来看，其源于应对自动化技术的威胁，"经自动化处理"可谓是个人信息早期的构成要素之一，自动化处理下的个人信息保护是《个人信息保护法》规制的重点(关于个人信息合规审计与算法审计之间的关系，下文将进一步进行讨论)。

这表明我国基本确立了对于算法的审计制度，因此，算法审计已然成为法律要求。目前，学界对于算法审计制度构建缺乏充分的讨论。算法审计有何制度价值？什么是算法审计？如何实施？成为亟待解决的法律课题。据此，从价值论出发，阐明设置算法审计制度的理由，并从算法审计对象、目的、作用范围出发厘清算法审计制度的内涵，最后，提出构建我国算法审计模式的制度路径。

一、算法审计的制度价值：设置算法审计制度的理据

算法审计有何制度价值？在论述算法审计的规则构建之前，有必要对"为什么要设立算法审计制度"进行规范分析，明确算法审计制度设立的现实意义。

(一)规范和监督算法权力，维护数据正义

算法审计是监督算法权力的运行、维护数字正义的有力武器。迈克尔·曼(Michael E.Mann)在《社会权力的来源》一书中指出，权力是通过支配人们的环境以追逐和达到目标的能力。网络社会由一个个智能平台系统组成，算法成为网络社会的"法律"，支配着人们的环境，并且能够通过这种力量影响网络"共同体"，从而达到自己的目的。近年来，越来越多的学者意识到算法对网络

世界的“支配”,以及这种支配力量会延展至现实社会,学者将这种力量命名为“算法权力”。信息革命是人类从物理世界向网络世界的一次“大迁移”,但数字社会中权力行使方是算法以及算法背后的操控者,数字权贵可以为了欲望和敛财罔顾他人尊严、自由甚至生命,我国就曾被爆出外卖平台滥用算法压榨骑手的事件。尤瓦尔·赫拉利(Yuval N.Harari)在其书《未来简史》中提到,由于人们无力处理海量数据,逐渐将权力交给算法,随着算法不断占领市场,财富和权力可能会集中向拥有强大算法的技术精英手中,造成前所未有的社会和政治不平等。算法异化将会对数字社会的正义带来极大的危害,有必要把算法权力关在法治的“笼子”里。

一般而言,为制约权力的肆意使用,法律会通过设置“问责制”(accountability)的方式来使得权力的拥有者正当行使职责,对公民负责。面对算法产生的权力,各国立法机关也提出了算法问责制(algorithmic accountability)以保障算法权力的正当行使。算法问责包括明确对算法的基本要求,并通过算法影响评估(algorithmic impact assessment)、算法审计(algorithmic audit)等活动实现监督,确定算法违法后果及救济途径等。学者将算法问责制分为:①原则和指南;②禁令和暂停;③公共透明度;④影响评估;⑤审计和监管检查;⑥外部/独立监督机构;⑦听证权和上诉权;⑧采购条件。如何实现算法问责成为需要解决的问题,欧盟的GDPR以及美国的《算法问责法案》均提到了透明度、认证等事前监督检查的方式,而算法审计无疑是最有效的监管措施之一。

正如为解决监督政府是否规范执行预算而实施政府审计一样,算法审计也是通过有胜任能力的专业人员对算法进行检测监督,及时发现算法中的问题,能够预防算法主体(本文的算法主体即指算法的设计者或者运行方,不包括消费者、用户)滥用算法,以确保算法权力被正当使用,维护数字社会的正义。

(二)防范数字风险,应对算法权力的弥散性

德国社会学家乌尔里希·贝克(Ulrich Beck)在其著作中提出风险社会的概念,科学带来文明的同时创造了风险,而且这种风险是全球性、系统性的风险,这种风险使得人类日益“生活在文明的火山口上”。风险社会与技术发展相伴而生,技术的不确定性对人类的生活乃至生存都产生巨大的威胁,因此技术的规制成为数字经济发展过程当中必须解决的难题。技术的规制需要按照找寻技术的特征,对症下药。

算法产生的风险源于算法权力的异化,这种异化表现为算法技术以不正当的价值观行使权力。因此对于算法的风险治理则需要把握算法行使权力的方式,与传统的政治学意义上自上而下的操纵和支配权力有所不同,算法权力通常呈现网络结构化和弥散性的特征。弥散性权力与权威性权力源自迈克尔·曼对权力类型的区分,与命令和服从的权威性权力不同,弥散性权力是一种“无所不在的权力”,以一种更加本能的、无意识的、分散的方式分布于整个人口之中。

算法通过隐蔽的、潜移默化的形式来实现权力的“运行”,导致被算法“支配”的用户难以直接感知到算法对其的危害。算法使用带来的信息茧房效应使得个人获得的信息往往是验证自己信念和观点的信息,不会出现“不同的声音”,这种情况下由于有限的、定制化的信息限缩了他们对世界的认知,进而影响个人的实际选择,而平台得以借此来影响个人的自主决定,例如剑桥分析事件。

越来越多的研究表明,算法被不断使用的过程不仅会侵犯个人尊严或自主权,还会加剧或扩大社会不平等。例如本·格林在其书中描述了城市技术导致社会和政治不平等的加剧,其与萨勒蒙·维尔约恩(Salome Viljoen)在论文当中也表明算法形式主义正在巩固不利的社会条件、歧视和不平等,索伦·巴洛克(Solon Barocas)和安德鲁·塞尔布斯特(Andrew D.Selbst)的研究表明,数据挖掘可能对受保护类别产生歧视性影响的机制。社交网络“情绪传染”实验表明平台可以通过信息推送的不同节点和内容来影响个人的情绪,2021年9月Facebook前首席产品经理弗朗西斯·豪根(Frances Haugen)向美国联邦执法局举报称:“Facebook自己的研究表明,它(算法)放大了仇恨、错误信息和政治动荡,但仍然选择隐瞒了这一研究结果。”在随后接受哥伦比亚广播公司“60分钟”节目采访时,豪根指出Facebook涉嫌利用算法掀起种族暴力、伤害青少年身心健康及鼓励分裂社会等多种“罪责”,在社会公益和自我利益面前选择了后者。因为信息不对称的原因,Facebook滥用算法权力的过程普通民众难以感受并知悉,甚至成为其中的助推器。

正如在所有者与管理者分离的现代企业中,所有者难以直接全面地监控管理者的行为时,会设计出财务审计制度一般,算法审计制度通过独立且专业的第三方审查方式。采取定期检查代码、测算运行结果等针对性的测试方式对算法是否异化作出评价,及时检测算法技术对社会公益产生的风险,保障社会的健康发展。

算法所产生的风险,除了上述提及的对就业公平、人格尊严等个人权益产生侵害的可能性之外,其作为人工智能系统的核心,还存在对人类整体安全产生威胁的风险。人工智能算法的算力远超人类,这将会导致人工智能对人类整体安全产生重大的不可控的威胁。物理学家、宇宙学家史蒂芬·霍金(Stephen W.Hawking)在生前的最后几年不断提出对于“人工智能”的担忧,他曾表

示:“人工智能的崛起是人类历史上最好的事情……然而人工智能在未来可能对人类生存带来毁灭性的威胁,除非我们学会如何规避、控制风险。”如何应对人工智能带来的威胁已然成为全球瞩目的话题,为规范人工智能的发展,许多平台组织和政府纷纷提出人工智能伦理规范。据不完全统计,迄今为止国际上由不同国家、机构和团体公布的人工智能和机器人伦理标准文件已多达50余份。我国科技部2021年9月26日发布了《新一代人工智能伦理规范》,对人工智能提出了增进人类福祉、促进公平公正、保护隐私安全、确保可控可信、强化责任担当、提升伦理素养等6项基本伦理要求。人工智能伦理规范频出,表明算法审计所注重的价值观并非简单的针对算法对于个人信息的保护,还包括对一些算法在人类整体安全等多个方面的价值观进行检测。

(三)控制算法异化,平衡算法透明要求与商业秘密保护

算法审计制度能够有效解决算法的不透明性困境。工业革命以来,技术工具理论一度占据政府和政策对于技术规制的主导地位,技术作为一种“中性”工具被民族国家战略性地鼓励研发、促进和应用。尽管技术工具理论存在不合理之处,但技术发展过程中仍然存在技术工具主义的影子。在技术工具理论中,技术被作为经济利益被赋予了合法的秘密性,即使技术异化也会因为秘密性而难以被监控和规制。算法的秘密性既表现为算法技术在输入和输出之间出现了常人难以了解和把握的“逻辑隐层”,如人们难以理解AlphaGo的棋局,也表现为对算法秘密的保护使得算法决策权被带进了不透明的区域,这为算法的异化提供了便利之门。算法秘密性将算法披上了“表面中立”的面纱并将“歧视本性”隐藏于“算法黑箱”背后,即使推动算法透明已然成为各国共识。例如,我国《个人信息保护法》第二十四条也作出了利用个人信息进行自动化决策时应当保障透明的规定,但算法的透明公开与算法技术的秘密性保护的平衡无疑是法律政策上的难点。

2021年10月,G7集体达成关于数字贸易原则的共识,其中明确要保持有效和平衡的知识产权框架,保护商业秘密。算法的秘密性使得常人难以获得算法的相关信息,算法主体拥有对算法的单方解释权,并且算法主体作出的解释对于不熟悉算法技术的普通人来讲也是诘屈聱牙,难以理解。算法审计以内部审计为原则,以专业机构的外部审计为补充,能够有效地克服以上问题,审计过程的保密性要求使得作为商业秘密的算法利益得以保护,同时,审计人员通过更为科学的方式(代码审计或者通过其他替代性的检测方式,如对算法输出结果进行检测)探析算法黑箱,站在更加专业的角度对算法是否异化进行分析,以第三方的立场解释算法、得出结论,节省了普通大众探索算法技术的成本。

可以预见,随着大数据和人工智能的广泛运用,算法审计将是规范算法权力,确保算法向善、合法、合乎道德和安全的重要制度选择。

二、算法审计的制度内涵:法律属性分析

制度的构建需要厘清制度的内涵。目前,国内外并未严格明确算法审计的定义,一些立法中甚至混淆使用算法影响评估、算法审计等制度,例如《算法问责法案》所指的算法影响评估就有算法审计之意。但两者在适用方式和目的上存在一定的区别,算法风险评估一般是指算法设计者自身对于未上市的算法进行风险评估(并可能在上市后持续评估),定期测试其所可能存在的风险;而算法审计强调通过政府或者第三方对于已经上市的算法进行监督。在我国,尽管人工智能发展的实践中存在对算法审计的需求,并且已经出现算法审计的立法规定,但对于算法审计的制度内涵仍没有统一的认识,导致算法审计活动的实施缺乏认知上的统一性、规范性。实际上,算法审计是基于合规需要而对算法开展的专业性监督活动,可从审计的对象、目的、作用范围出发去把握该制度的本质。

(一)技术审计:算法审计的审查对象分析

审计活动的对象是对审计内容所作的理论概括,算法审计的对象即回答需要通过检查什么内容来检测算法是否符合价值规范要求,是否存在数字风险。从发展历史来看,早期的算法审计主要关注数据的收集和系统开发模型,重点考虑数据样本、系统的准确性与计算成本。在机器学习(ML)算法普及之后,数据和模型的交互关系较以往变得更复杂,算法审计相应从关注算法的性能、效率转为主要围绕偏见、可解释性和数据安全保护。可见,虽然审计过程关注的规范内容会发生变化,但算法审计活动的对象仍然主要是算法模型、数据以及相关活动。

1.算法模型是算法审计的首要对象和核心内容

算法模型的设计是企业操控算法压榨弱势群体的直接方式。学者指出目标函数(或目标)不正确或不精确的AI可能会以不理想的方式行事,误用及设计不当都有可能产生风险。2021年初,英国竞争与市场监督管理局(CMA)就总结了平台通过算法操控侵害消费者权益的各种形式,包括个性定价(数据杀熟)、个性排名、操纵用户行踪、算法歧视等。算法设计者通过有偏向性的“回报函数”不断追逐其价值取向,获取不正当利益。

值得注意的是,目前被广泛运用的深度学习算法,还可能出现算法因自适应学习也产生异化,例如微软的机器人聊天程序Tay就曾被质疑在半天内从Twitter上的用户习得仇视人类和种族歧视。而随着系统智能化的逐步提高,AI无法预料的行为将变得越来越危险和难以

的业绩；合规审计则是查明被审计对象是否符合法律、合同或有关控制标准而实施审计活动。就算法审计而言，其目的有别于财务审计和经营审计，并非为了监督算法的效率性、经济性，而是为了对算法是否合乎法律法规、社会道德等的价值规范要求进行鉴定、分析，是法律在算法技术理性的规制监督和规范措施，即算法审计属于合规审计。事实上，我国《个人信息保护法》第五十四条、第六十四条也明确了算法审计的合规属性。尽管新近发布的《互联网平台落实主体责任指南(征求意见稿)》第八条、《网络数据安全管理条例(征求意见稿)》第五十八条都提出了安全审计的概念，但从其规范表达的内容来看，实质上仍然体现了合规审计的本质属性。

(三)风险审计：算法审计的适用范围分析

算法成为人工智能时代的新型社会风险源。如何预测、识别算法违规的风险，评估其程度和所带来的损失，防范风险累积或者传递并转化为现实的危害，控制和化解算法风险是算法审计必须直面的问题。这些问题，决定了算法审计的规范范围为一切可能引起风险和现实安全危害的算法。保障安全是AI基本伦理价值之一，这一点在各国政府、国际机构和团体组织的伦理要求中都有所体现，如美国未来生命研究所(FLI)发布的《Asilomar AI原则》、G20发布的AI准则、迪拜政府公布的《迪拜AI原则》、加拿大财政委员会秘书处(TBS)公布的《政府使用AI系统的七项原则》。我国国家新一代人工智能治理专业委员会2021年颁布的《新一代人工智能伦理规范》(第一条及第三条的有关内容)中就明确指出促进人工智能公平、公正、和谐、安全是基本伦理要求。因此，2021年发布的《互联网平台落实主体责任指南(征求意见稿)》《网络数据安全管理条例(征求意见稿)》等规范都进一步强调算法的安全审计取向。从这个意义来看，算法审计具有风险导向型审计的重要属性。

所有的算法都是值得审计的。诚然，随着数字经济的发展，各行各业都开始数字化转型，大至政府、平台型互联网企业，小至初创公司、个人，都有可能会设计或运用算法以提高自身经营或管理效率。从风险的全覆盖以及合规遵从的普遍性来看，有必要对所有的算法都要求进行内部的合规审计。

从目前各国的立法实践来看，考虑到制度成本以及审计成本，许多国家并未要求所有的算法都进行外部审计。例如加拿大(加拿大《自动化决策指令》第6条，并且其以附录的形式将风险等级以及相关的监管要求予以详细说明，详可见其附录B和C)根据算法对个人或群体权利、健康福祉、经济利益、生态的可持续性等因素影响级别分为没有影响、适度影响、重大影响、非常大的影响四个等级，对于没有影响的算法不需要同行评审，具有适度影响、重大影响的算法要求至少一个同行评审，而非常大的影响则至少两个同行评审。欧盟数据保护影响评估制度(DPIA)是基于风险的保护路径(risk-based approach)评估数据处理活动的必要性和适当性的问责制工具，该制度将风险作为是否展开评估活动的基本衡量指标，欧盟出台的《关于数据保护影响评估和确定数据处理活动是否"可能导致高风险"的指南》中对什么是风险亦进行了说明。GDPR第35条"数据保护影响评估"以及相关说明第91条"关于数据保护影响评估的必要性"指出，并非对所有数据处理活动进行DPIA，只有可能对自然人权利和自由造成高风险的情形下，DPIA才会成为强制性义务。基于风险的保护路径在《人工智能法》中也得以体现，该法按照风险等级将人工智能分为"最低、有限、高、不可接受"四个等级，对不同风险等级人工智能系统设定了不同级别的限制。

我国立法部分吸收借鉴了此种风险分级治理理念，在《个人信息保护法》出台前夕，全国信息安全标准化技术委员会牵头起草的《信息安全技术 个人信息安全影响评估指南》规定"个人信息处理活动自身可能涉及对个人权益影响及相应风险较高的情况下，需开展个人信息安全影响评估"，并在附录B中对"可能导致高风险"的个人信息处理活动作列举说明。《个人信息保护法》则对具有较大风险或发生安全事件的算法，规定了接受审计的义务，将风险导向型管理理念引入个人信息合规审计领域。在平台算法的规制方面，市场监管总局发布的《互联网平台分类分级指南(征求意见稿)》第三章对于容易引发并传播风险的大型互联网平台，依据"用户规模、业务种类、经济体量、限制能力"等四个因素将平台分级为超级平台、大型平台及中小平台，并在随之颁布的《互联网平台落实主体责任指南(征求意见稿)》第八条中规定了属于大型平台以上等级的平台需要进行审计。

算法审计制度亦有必要需依据算法风险等级(风险等级的考量因素可依据以上所提及的相关法律的规定以及借鉴域外确定风险级的有关制度予以设计)设计为四个等级，按照不同风险等级的算法设计不同的审计要求(详见表1)。

三、算法审计的制度实现：我国模式的构建

根据上文的分析，我们可以总结出算法审计的基本定义，即算法审计是针对被审计单位的算法模型、数据及有关技术活动的合规性、风险性进行审计，用以监督算法的正当使用，推动智能技术向善，保障国民经济和社会健康发展的活动。而构建算法审计制度并将制度定型化，则需要讨论算法审计模式，主要涉及由谁作为审计主体，通过哪些制度规则和程序机制来执行算法审计活动，从而将算法审计制度在现实社会生活发挥作

表1 算法分级审计问责框架

审计形式 / 风险等级	内部审计	外部审计
一级	无需审计	无需审计
二级	定期进行内部审计，并保存审计底稿、日志等记录以备检查	发生安全事件等特殊情况下方需进行外部审计
三级	定期进行内部审计，并对外公布内部审计报告	发生安全事件等特殊情况下方需进行外部审计
四级	定期进行内部审计	定期进行外部审计，并对外公布外部审计报告

用。从运行的角度来看，算法审计模式必须包括实施主体、实施条件、实施机制和实施后果四大方面。

纵观域外的立法状况，许多国家(地区)均针对算法出台专门的立法予以规范，例如美国的《算法问责法案》《算法公平法案》、欧盟的《人工智能法案》、加拿大的《自动化决策指令》等。目前，我国还缺少专门针对算法进行规范的法律，对于算法的规制散落在各部相关的法律法规之中。其中法律如《个人信息保护法》《数据安全法》《网络安全法》《电子商务法》等，法规如《互联网信息服务算法推荐管理规定》《平台责任指南》《关于加强新时代文艺评论工作的指导意见》《关于维护新就业形态劳动者劳动保障权益的指导意见》《互联网信息服务算法推荐管理规定(征求意见稿)》等。这导致对于算法治理“各自为政”，缺乏体系化，仅仅通过《个人信息保护法》《数据安全法》等法难以实现对算法的全面规制。

鉴于算法对于社会发展的重要程度，建议出台《算法规制法》，将算法审计纳入算法问责框架，通过专门的立法确立算法审计制度的目的、范围、对象，并对算法审计的实施主体、实施条件、实施机制、实施后果进行规定，将算法审计制度确立为算法监督的法定方式。

(一)算法审计的实施主体：审计体制的确立

算法审计的实施主体，实践中可分为内部审计主体与外部审计主体。不论是内部审计还是外部审计，都必须考虑算法审计主体从事技术、合规和风险审查上的主体适格性，避免出现“外行指挥内行”的问题出现。

根据《个人信息保护法》第55条的规定，个人信息处理者需要针对算法进行内部审计。内部审计是一种推进和完善自我治理的方法，可以由算法的设计、运营和使用的主体通过建立人工智能内部审计机制经常性地检查评估算法的合规性与风险状况。有条件或者规模较大、人工智能运用较为广泛和深入的组织可以在建立和健全法制、合规部门时考虑设立独立的、专门的算法审计机构，一般的单位可以由法制部门、合规机构联合技术机构去实施算法审计。需要审计的主体也可以自行委托适格的社会独立机构帮助其开展内部审计。

何为适格的独立算法审计机构？目前，我国还没有具体的制度安排，未来可能的制度选择有三：

一是将算法审计纳入财务审计的业务之中。会计师事务所传统的审计业务主要是针对财务报表及其附注是否存在错报、是否为公允反映而发表专业意见，如今也开始涉及对网络信息系统的安全性进行审计(内部控制审计活动中的一环)。但算法审计与财务审计在对象、内容和目的要求等方面明显不同，会计师事务所是否有能力胜任算法审计还有待考察，因此，直接将算法审计活动纳入会计师事务所财务审计活动当中缺乏科学合理性。

二是由公共审计机构进行审计。挪威、芬兰、德国、荷兰以及英国目前都通过公共审计的方式对公共部门算法进行审计。我国虽然存在财经法纪审计这种合规性审计，但主要是对政府机关、事业单位和国有企业而开展的具有外部强制性的经济监督工作，审计的目的是检查监督有关单位是否存在财经性质违法行为，技术安全合规并不在传统公共审计的范围之内。同时，公共审计所针对的对象主要是公共部门的资金，并不针对私营单位，公共审计部门实施算法审计容易混淆算法审计的职能部门与经济监督主管部门之间的职责关系，因此公共审计机构显然不适合作为一般企业审计的委托机构。

三是将算法审计作为一种全新的民间审计，由专门的算法审计组织对此进行审计，例如美国的奥尼尔风险咨询与算法审计公司(ORCAA)就是一家专门针对算法进行审计的企业。这种做法主要是考虑到算法审计存在高度专业性，且审计对象与财务报表存在弱相关性，不属于注册会计师审计的业务范畴，应当由算法领域的权威人士专门进行审计。《个人信息保护法》第64条规定，履行个人信息保护职责的部门对个人信息处理活动存在较大风险或者发生个人信息安全事件的算法，有权要求个人信息处理者委托专业机构实施算法审计。由专门的算法审计机构开展专门的算法审计将是未来的主流选择，国家需要及时出台准入的门槛要求和审计活动的规范标准。

通过以上分析，以市场主体来解决算法伦理问题，可能更加符合我国当前所推崇的市场导向的要求，因此，将算法审计定位为一项全新的民间审计活动更为恰当。鉴于当前算法审计经验仍有待探索，尚无法完全确定适格的审计主体，律师事务所、计算机研究机构、会计师事务所等均具有从事该项活动的潜质，可以采用审批认证的方式确立算法审计主体。即有意从事算法审计的单位组织可以向有关机关提出申请，有关机关根据对申请人资质的考察，决定是否认证其为算法审计适格主

体,并认可其从事算法审计业务。

鉴于《个人信息保护法》及《数据安全法》均认定网信部门作为相关治理活动的统筹协调部门,并且根据《互联网信息服务算法推荐管理规定》所设立的算法备案系统亦由网信部门进行管理。无论从法律体系性还是技术治理经验方面考量,网信部门统筹负责算法规制活动较为合适。因此,《算法规制法》应将网信部门作为主要责任机关,同时规定算法审计组织的认证工作、审计基础的制定工作、算法问责工作等由网信部门主持。

(二)算法审计的实施条件:基本要求的准则化

传统审计是客观地获取和评价与经济活动和经济事项的认定有关的证据,以确定这些认定与既定标准之间的符合程度,并把审计结果传达给有利害关系的用户的系统过程。而这个过程顺利实施的前提要件之一,是要存在相应的参照标准。注册会计师在财务审计过程中的参照标准,一般是指企业所依据的会计准则。

要评价算法的规范行为,必须描述(最好是正式指定)反映先前确定的规范的标准行为,描述用户期望算法做什么,验证算法是否真的这样做,并验证规范是否真的符合标准。目前,我国相关法律关于算法的要求都过于原则化,缺乏可执行性,尤其是针对价值观的规定缺乏确定性,难以直接作为参照标准。以"公平"为例,不同的人对于公平的定义是不一致的。美国非营利媒体"为了人民"(ProPublica)对"替代性制裁的惩罚性罪犯管理分析系统"(Correctional Offender Management Profiling for Alternative Sanctions,COMPAS)的审计结果与开发者的审计结果大相径庭,双方在缺少具体明确的解释下对"公平"的理解明显不同。因此有必要根据基本要求设置一定的标准,这既为算法主体设计、训练和运行算法提供指引,同时也为审计师进行算法审计提供参照系。

算法审计既包括对于数据的审计,也包括对于算法的审计。从数据的合规审计来看,可以依据《个人信息保护法》《数据安全法》等进行解释和制定,我国目前已经存在一些数据合规指引。例如,国家市场监督管理总局、国家标准化管理委员会联合发布的《信息安全技术个人信息安全规范》(2020年版),中国人民银行发布的《中华人民共和国金融行业标准〈人工智能算法金融应用评价规范>》等。再如《征信业管理条例》《电信和互联网用户个人信息保护规定》《全国人民代表大会常务委员会关于加强网络信息的决定》《网络安全审查办法》《App违法违规收集使用个人信息自评估指南》《网络安全等级保护基本要求》《大数据安全管理指南》等,都存在关于数据合规相关规则。

从对算法价值观审计来看,应当设立算法价值观标准,针对设计算法及算法运行过程能做什么、不能做什么作出行为指引,这些指引既包括对于算法设计者、控制者的行为指引,也包括对算法的"行为"作出指引。例如,设计算法时不得存在偏向的回报函数,尽量防止算法的过度拟合(算法不能很好地泛化到新数据)或过度简化(也称欠拟合,是指算法没有很好地描述数据),算法做出的决策应当维护人类的基本安全等。价值观标准应当是大众都普遍接受的价值观,是社会所认同的"共同善"。

算法的价值观是算法作出行为的指引,其应当以符合特定区域的方式行事,即算法的行为应当反映一国(地区)社会的价值观。从各国关于人工智能伦理的有关政策来看,皆提及要体现和保护当地的价值观。例如,欧洲科学和新技术伦理小组发布的《关于人工智能、机器人和自主系统的声明》(*Statement on Artificial Intelligence, Robotics and Autonomous Systems*)中,就提及系统需遵循"《欧盟条约》和《欧盟基本权利宪章》规定的基本价值观",美国发布的《关于保持美国在人工智能领域领先地位的行政命令》(*Administration Executive Order on Maintaining American Leadership in Artificial Intelligence*)中,也提出在人工智能应用中保护"公民自由、隐私和美国价值观"。

因此,我国的算法价值观合规标准,应在联合国《人工智能伦理问题建议书》呼吁的尊重人权等基本价值观基础上,结合我国社会主义核心价值观进行制定。为了增加价值观的可落实性和适应性,还需要结合算法所应用的行业特征对价值观标准予以细化,以义务或禁止等更容易转化为计算机语言的表达方式设计出适用于各行业的实施细则。由于算法的发展日新月异,价值观标准及实施细则以实践中可能出现的风险为导向进行设计,并根据实践新出现的问题对价值观标准进行不断补充完善。目前,主要是针对商业企业不能实施大数据杀熟、算法偏见等问题,如何设计、训练和运行算法,涉及何种情形将会认定为大数据杀熟、算法偏见,再如无人驾驶汽车内载算法应当遵循维护驾驶员和第三人生命的原则进行设计。

同时,由于算法和人工智能的开发和应用以企业为主导,因此有必要通过制定有关的企业运行流程以确保企业真正落实数据合规及价值观有关指引。目前,在我国的内部控制规范体系中,主要关注与财物及战略经营相关的控制,缺少与数据、算法相关的内部控制指引。有必要考虑设计相关的内部控制指引,指导企业建立相关的内部控制帮助企业实现针对性的管理。例如,设计被审计单位在算法设计、训练和运行时的审计追踪(audit trails)机制,设计系统操作步骤的可追溯日志,以便检查过程中的不当行为及后续的问责,在美国《算法公

平法案》第五节“公平程序”中就规定保留至少5年的审计跟踪记录，设计算法道德设计师岗位，算法的开发、修改需要经过算法道德设计师进行验证签字后方可进行，以及设置内部算法审计架构的指引等。

(三)算法审计的实施机制：具有针对性和可行性的可审计性建构

算法的可审计性是指通过披露能够进行监控、检查或批评的信息，包括通过提供详细的文档、技术上合适的API和宽松的使用条款，使第三方能够探查、理解和审查算法的行为。在财务审计过程中，被审计单位的配合义务包括：管理层已认可并理解其对财务报表承担的责任，包括编制财务报表的责任，设计、执行和维护必要的内部控制，以及为注册会计师提供必要工作条件。因此，可审计性涉及审计活动的第二个前提要件，即被审计单位配合审计师进行审计。在《人工智能伦理问题建议书》第43、53、56条中，联合国教科文组织就建议会员国确立和提升对人工智能的可审计性。

从算法审计的过程方法来看，除了传统的注册会计师所使用的审计程序，例如询问、观察、检查、重新执行等，实践还出现了许多专门针对算法审计的研究方法。代码审计(code audit)，即针对算法的代码进行审计，检查是否存在不当的设计。但代码审计涉及秘密性问题，一些算法的机制使得算法难以公开其代码，例如Reddit软件为了防止垃圾邮件机器人使用公开的算法攻击Reddit，使其评级和评论系统无用，该算法的内核(称为“投票模糊”代码)必须保持闭源和秘密，并且由于算法自学习过程也会导致价值感异化，往往仅分析代码难以完全得出算法价值观正当的结论。

鉴于此，实践中还设计出了针对算法输出结果的审计方法：一是非侵入式用户审计(noninvasive user audit)，即与算法的用户进行沟通，询问用户使用体验或者取得用户同意后查看其账户、检查其使用算法的痕迹等，类似于审计的询证程序；二是抓取审计(scraping audit)，也称爬虫审计，是指审计师通过API直接访问算法，对算法的输出结果进行任意抓取，但抓取行为可能涉及对平台(算法)的侵权问题，并且目前许多平台也启用了反抓取机制；三是虚拟用户审计(sock puppet audit)，审计师利用计算机程序来冒充用户进行审计，虚拟用户涉及欺骗的问题，平台可以此为由主张用户数据存在问题才导致算法偏见；四是众包审计/协作审计(crowdsourced audit/collaborative audit)，指审计师招募足够多的“测试人员”(用户)群体对算法进行测试，此种审计方法成本较大，而且对用户较多的平台才有效果。

不难看出，以上审计方式都存在可审计性的问题，如代码审计需要取得被审计单位同意才能获取其代码，非侵入式用户审计需要被审计单位提供用户信息，抓取审计需要取得被审计单位访问算法API的许可，虚拟用户审计需要事前取得被审计单位对虚拟身份的认可，众包审计则涉及审计成本负担问题。

因此，应在《算法规制法》中考虑以上问题，明确可审计机制，设置被审计单位配合进行算法审计的义务，加强算法审计的可审计性。同时，由于算法审计的专业性和复杂性，有学者亦提出可以考虑由算法审计的公共部门依据算法审计的要求，设计相应的用以测试算法价值观的算法。南加州大学研究人员研发的DeepTrust工具可以对AI算法生成的数据和预测的信息进行验证；日益成熟的“歧视感知数据挖掘“也能够识别偏离公平伦理的算法，通过技术进路”以算法控制算法”从而实现算法可信正在成为可能。芝加哥大学就开发了审计工具“埃奎塔斯”(Aequitas)，可供机器学习开发人员、分析师和决策者审计机器学习模型中的偏差。

(四)算法审计的实施后果：审计结果问责的确立

上文提及，算法问责制是实现算法权力监督的重要手段，而算法审计则是算法问责中的重要一环，但如何通过算法审计实现算法问责呢？需要通过制度赋予算法审计以法律效果，将算法审计作为问责框架中的重要一环。

其一，在检测出算法异化时，审计师有权要求被审计单位对此进行说明并加以改正，如果被审计单位在经沟通后仍然拒绝改正的，审计师应当出具“否定意见”的审计报告并向有关机关报告审计结果，请求有关机构对相关问题进行处理。

其二，审计报告可以作为有关机关进行行政处罚的依据。我国《个人信息保护法》第七章、《数据安全法》第六章均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，而前文提及算法审计包含对于数据的审查，因此，算法审计的结果可以作为行政机关依据以上法律对数据违规行为作出行政处罚的依据。我国目前因为不存在《算法规制法》，并未专门针对算法设置违法责任。而诸如《人工智能法》规定设置禁止类AI(第五条)，或者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，为算法设置专门的“行政处罚”。因此，应当在《算法规制法》中设置对于算法的特殊责任形式，除了常规的行政罚款之外，也有必要为算法设置一些特殊的责任方式，如禁用、暂停或者召回等。《算法规制法》的责任条款将会作为算法违规时的处罚依据，以此实现对算法的问责。

其三，算法审计的结果可以作为被侵权人提起侵权

（下转第10页）

论政治责任审计的基本概念框架

——基于党内巡视的视角

熊宇昊

党中央把完善党和国家监督体系作为推进国家治理体系和治理能力现代化的重要举措。为了更好将党和国家监督体系的制度优势转化为治理效能，有学者从贯通国家审计监督与党内监督出发，提出可以发展政治责任审计，构建政治责任审计与经济责任审计双轨并行的综否审计模式。党内巡视是中国共产党进行党和国家自我监督的有力抓手，而基于党内巡视的政治责任审计，则可成为中国共产党根植于中国传统文化，在新时代进行治国理政的一个自发性创新。

然而，当前关于政治责任审计这一概念的研究还处于空白，其概念框架还有待于进一步探索。本文基于党内巡视的视角，立足党内巡视工作的具体实践，尝试对政治责任审计的概念和特点进行界定与归纳，对政治责任审计的本质、指导思想、动力机制、主体、客体、内容、目标、方法、依据规范和制度环境进行概括阐述，以期初步形成政治责任审计的基本概念框架。

一、政治责任审计概念形成的可行性

不同学者对审计定义的表述有所差异。秦荣生(2017)把审计的概念界定为：独立客观的经济监督、确认和鉴证活动。美国会计学会(AAA)(1973)将审计定义为："审计是通过客观地获取和评价有关经济活动与经济事项认定的证据，以查明这些认定与既定标准之间的符合程度并将其传达给利害关系人的一个系统过程。"这一定义侧重于信息的传递。陈汉文将其提炼为四个要素：胜任的独立人员、经济活动和经济事项的认定、认定与既定标准的符合程度以及客观的证据。上述定义都是从经济管理的视角对审计进行定义的。

现代审计理论认为，"审计具有两项职责，两者均与证据有密切关系，一是证据收集；二是证据评价"。美国会计学会前会长阿尔文·A.阿伦斯对审计进行了更为宽

(上接第9页)

诉讼、有关组织提起公益诉讼的依据。算法歧视、大数据杀熟等问题都将导致用户的权益受到实质损害，算法审计报告的主要预期使用者是行政机关及使用过算法的用户，自然可以依据算法审计报告的内容对算法主体提起侵权诉讼。同理，我国《个人信息保护法》规定了个人信息保护公益诉讼，在个人信息遭遇侵权时可以由人民检察院、法律规定的消费者组织和由国家网信部门确定的组织提供公益诉讼，算法审计结果亦可以作为起诉依据。

四、结语

数字社会的治理不仅仅是对人的治理，还是对数字技术的治理。算法作为数字社会的核心要素，被不当使用将为个人及社会带来巨大隐患。正如福柯在《规训与惩罚》一书中提到，现代国家不再对人的生理肉体施行其权力，而是将权力逐渐转向了心理层面。人们开始受到新的控制机制，通过律条规定对人的精神进行"驯服"。德勒兹在福柯的基础上，进一步提出这种律令的控制正在逐渐演变成为机器的控制，人们开始被算法"驯服"。

本文基于实现对算法的治理、引导算法向善的目的，对算法审计制度进行了探讨。当前尽管《个人信息保护法》提及了个人信息审计，但并未就具体实施方式作出规定，并且个人信息审计难以囊括算法审计的所有要素。因此，有必要设立《算法规制法》，将算法审计明确为法定算法规制方式，对算法审计的实施主体、实施条件、实施机制、实施后果进行规定，以便更好地指引算法审计活动。但本文仅在理论和整体制度设计方面提出了看法，对于如何制定算法审计的实施细则，仍有待审计学、法学、计算机科学等多个学科领域共同合作，群策群力进行设计，推动算法在法治轨道上实现善治。

基金项目：科技部国家重点研发计划"跨部门跨地域社会信用治理关键技术研究与应用示范"(项目编号：2022YFC33032C00)；国家哲学社会科学基金重大专项"社会主义核心价值观融入网络治理法治化研究"(项目编号：19VHJ005)。

作者简介：张永忠，华南师范大学法学院教授，研究方向为经济法与数字法学；张宝山，暨南大学法学院民商法博士研究生，研究方向为民商法与数字法学。

原载《电子政务》(京)，2022.10.48～61