# 算法规制的路径转向:基于元规制理论的算法审计*

杨永兴

(中国海洋大学 法学院,山东 青岛 266100)

**摘　要**:算法技术加速演变,呼唤算法规制迭代升级。基于敏捷治理、场景规制、适应型治理、回应型规制等理论衍生出的个体赋权、算法透明、“以技治技”等传统算法规制工具不仅引发了规制主体与规制受体之间的冲突,还难以因应生成式人工智能算法带来的规制挑战。而基于元规制理论开展的算法审计作为一种反身性动力机制,可以有效弥补传统算法规制工具的功能缺陷,协调算法规制范畴内各方利益冲突,并有助于破解算法安全与算法发展的统筹治理困境。因此,元规制理论指导下的算法审计应成为我国算法规制发展的应然方向。在参考域外立法经验的基础上,结合传统审计学领域的审计实践,未来我国算法审计的规范构造可考虑围绕算法审计的审计原则、审计主体、审计方法、审计内容、审计效力等几个方面具体展开。

**关键词**:算法;算法规制;算法审计;算法规范

**中图分类号**:TP18;F239.1;D922.16　**文献标识码**:A　**文章编号**:1672-335X(2025)05-0104-12

**DOI**:10.16497/j.cnki.1672-335X.202505010

## 一、引言

自第四次信息科技革命以来,以算法为代表的新兴数字技术悄无声息地嵌入社会生活的各个场域中,通过无所不在且难为人知的方式重塑人类社会生存环境,引领社会生产变革,拓展国家治理疆域,甚至催生人类文明新形态。[1]与此同时,作为算法技术创新赋能人类社会千行百业、增进人类脑力、提高社会生产力的代价,算法应用所带来的算法霸权、算法剥削、算法歧视等“创造性破坏”正不断挑战人类社会的基本价值共识,破坏社会基本秩序,侵蚀基本权利。

为回应算法社会化过程中的“创造性破坏”,国内外法学界相继提出敏捷治理、场景规制、适应型治理、回应型规制等算法规制理论,并基于这些规制理论发展出个体赋权、算法透明、“以技治技”等具体算法规制工具。① 不可否认,这些既有算法规制理论和规制工具在防范算法风险、引导算法向善、推动算法规范性文件的出台等方面发挥了不小的作用。但是,2022年末以来,生成式人工智能的兴起缩短了算法技术的迭代周期,其普适化应用扩散了算法技术的不确定风险,[2]进而对算法治理提出多方面的结构性挑战。这些结构性挑战致使传统算法规制理论及规制工具应对乏力。为充分因应生成式人工智能算法所引致的负外部性问题,算法规制范式亟须更新。

作为对后现代社会算法规制需求升级的回应,各国政策制定者与学者开始将目光聚焦至算法审计,

* **收稿日期**:2024-12-30

**基金项目**:国家社会科学基金一般项目“智能社会中国家信息能力变迁推进治理体系转型研究”(23BFX021);国家社会科学基金重大项目“当代中国数字法学基本范畴体系研究”(23&ZD153);山东省自然科学基金青年面上专项“人工智能时代的法律语言认知机制:实证研究及前沿应用”(ZR2024QC235)

**作者简介**:杨永兴(1999- ),男,河南驻马店人,中国海洋大学法学院博士研究生,专业方向为宪法学与行政法学。

①代表性研究参见:于文轩,刘丽红.算法规制的敏捷治理[J].新视野,2022,(3):66-72;周翔.算法规制如何场景化[J].东方法学,2024,(2):136-150;张凌寒.算法权力的兴起、异化及法律规制[J].法商研究,2019,36(4):63-75。

希冀通过算法审计驯服算法这一可能的“脱缰野马”。在政策制定层面上，加拿大于2019年针对政府部门算法自动化决策发布《自动化决策指令》，依据决策对个人或社区的权利、个人或社区的健康和福祉、个人或社区的经济利益、生态系统的可持续性的影响程度将自动化决策划分为四个等级，并针对不同等级作出不同的同行审计规定。2020年，芬兰、德国、荷兰、挪威及英国五国最高审计机关联合发布白皮书，为审计人员开展算法审计活动提供基本指引。2021年，我国颁布《中华人民共和国个人信息保护法》（以下简称《个保法》），其第54条与第64条对个人信息保护合规审计作出了原则性规定；2025年2月，国家互联网信息办公室发布《个人信息保护合规审计管理办法》，在《个保法》的基础上对个人信息保护合规审计的开展进行具体规定。② 在学理研究层面上，审计学领域已有学者就算法审计内容、审计方法等进行探讨，法学界也有学者开始着手对算法审计制度进行法治化建构。作为算法治理实践中的一类新兴治理工具，算法审计为后现代社会规制算法风险提供了新的视角和范式，同时也有望推动算法法律规制体系的升级。但是，目前全球范围内尚未形成一套完备的、体系化的算法审计制度，学界对算法审计的理论建构与规范构造也缺乏充分的讨论。既有算法审计研究的研究视野较窄，忽视了算法审计本身的交叉学科属性，未能充分回应算法审计的正当性等一系列核心问题。同时，即使有研究对算法审计的具体内容作出阐述，也存在着构造内容不完全、不充分以及可操作性不强等问题。有鉴于此，本研究将基于回应既有算法审计研究不足的问题意识，分析反思既有算法规制理论与算法规制工具的缺陷，证立算法审计的理论基础与正当性，构造算法审计的具体内容，探索提升算法审计可操作性的路径，以期对我国算法法律规制体系的完善有所裨益，并推动算法向善。

## 二、对传统算法规制理论与工具的审思

为应对算法技术快速迭代过程中的负外部性，有关算法规制理论、规制工具的研究和实践在当代法学界如雨后春笋般涌现，并逐渐衍生出场景规制、敏捷治理、回应型规制、适应型治理等算法规制理论及基于这些规制理论提出的算法透明、个体赋权、“以技治技”等算法规制工具的学说。

### （一）算法规制理论刍议

在对算法等新兴技术的监管中，由于政府的家长制作风在互联网技术治理领域根深蒂固，[3](P104)“命令—控制”型的科层式规制理念早期盛行。但是“命令—控制”型的科层式规制范式囿于政府与市场之间的信息不对称，使政府缺乏“形成快捷决策所依赖的具有信用品意义的符号”，[4]进而导致此种规制范式在管控数字技术风险时面临着规制效果不佳的现实困境。出于对“命令—控制”型规制理论的反思，学界开启了对更优算法规制理论的追求，敏捷治理、场景规制、回应型规制、适应型治理等规制理论纷纷涌现。

#### 1.敏捷治理

敏捷治理起源于20世纪末软件开发领域的“敏捷开发”。根据2018年世界经济论坛发布的《敏捷治理：第四次工业革命时代政策制定的重构》白皮书，敏捷治理是指“一套具有柔韧性、灵活性、流动性的行动，且能够促进适应性、以人为本的决策过程”。[5]据此可见，敏捷治理是针对发展迅速、影响广泛的新事物、新领域提出的一种治理理论，参与治理的主体广泛性与政策制定的时间灵敏性是其典型特征。敏捷治理特别强调治理工具和治理方式的灵活多变，[6]以便在问题发生之前预判问题的样态、类型以及可

②这表明，我国在制定法层面上基本确立了对于算法审计的需求，构建了算法审计的底层规则，使算法审计的规范构造具备合法性基础。虽然二者都具备监督、鉴证和评估的功能，但是个人信息保护合规审计并不能与算法审计等同，因为二者的作用范围与审查侧重点不同，且一般来说后者包括前者。在作用范围层面上，个人信息保护合规审计主要针对个人信息处理行为，而算法审计不仅要审计数据维度的处理行为，还要审计模型架构、外部权力介入等维度的行为。在审查侧重点层面上，个人信息保护合规审计主要关注个人信息处理的合法性，而算法审计不仅关注算法的合法性，还关注算法的合理性。有关个人信息保护合规审计的监督、鉴证与评估功能的研究参见：赵精武.个人信息保护合规审计的基本框架与制度衔接[J].法律科学（西北政法大学学报），2025，43(3)：83-97。

能产生的危害后果。理想中的敏捷治理不仅需要投入大量的治理资源,还需要驱动制度变革和技术应用的双轮机制。[7]但在实践中,技术驱动的敏捷治理不仅容易“俘获”政府,还容易让政府在面对算法治理时走向“共同无知”的困境,[8]由此导致驱动敏捷治理的制度变革机制面临政治势能难以为继的难题。

2.场景规制

场景规制是基于场景理论展开的算法规制。其核心要义在于不同场景下算法的使用主体、可能产生的风险、受到的规范羁束均有所不同,为此,应当根据现实情况差异为不同场景下的算法应用确立不同的规则。场景理论的适用前提理应是不同场景界限相对明确且可以划分,但伴随着新兴技术的推陈出新,重叠场景不断增多,从而对场景的划分与判断提出严峻挑战。除此之外,将作为社会技术系统的算法拆分为各个子系统,[9]寻求各个子系统的应对策略,但忽视作为整体系统的算法所需的一般性规则的规制思路只会陷入“只见树木,不见森林”的困境。

3.回应型规制

回应型规制深受塞尔兹尼克回应型法律理念的影响,是由澳大利亚法学家艾尔斯等人提出的一种“投桃报李/以眼还眼”式的规制理论。该理论将规制活动预设在互动式的场景下,强调根据规制受体的守法程度,采取警告、警示、制裁等对应的规制策略和方式。[10]后果的动态回应性是回应型规制最突出的特点,也正因如此,回应型规制往往遵循的是实害救济的矫正正义逻辑,仅着眼于规制主体对算法主体的责任追究,忽视了个案正义意义上的预防和救济。不仅如此,规制所承载的回应型法律只是对技术的制度性安排,不仅难以因应新兴科技的发展,还缺少道德维度和价值维度的分析框架,以致规制效果总不尽人意。

4.适应型治理

适应型治理由生态学韧性理论和自组织理论发展而来,是一种探讨如何应对复杂社会生态技术系统挑战的理论。其核心思想在于根据外部环境的变化动态调整治理策略。该理论认为,传统“命令—控制”型规制和纯粹的市场机制并非最佳规制手段,在面对治理复杂性较高的领域时,政府应当支持行业自组织的协作并保持治理举措的灵活性。适应型治理在新兴技术领域得到了较多的应用,但根据适应型治理概念的提出者——美国学者迪茨等人的观点,适应型治理成功的条件在于获取充足的信息、解决冲突、促使遵守规则、提供基础结构以及做好变革准备。[11]然而在实践操作中,适应型治理受到算法规制范畴内各方的利益冲突、企业与政府之间的信息不对称、企业与机器学习算法之间的信息不对称等复杂因素的影响,不仅不具备成功的条件,还面临实施难度较大的困境。

(二)算法规制工具反思

为实现对算法的规训,学界以及各国政策制定者基于前述规制理论提出赋予数据主体“权利束”、公开算法源代码、实施算法备案、强化技术治理等多种规制工具。本研究从抽象维度将这些规制工具大体分为个体赋权、算法透明、“以技治技”三类。每一类规制工具都正在算法规制实践中发挥作用且仍将长期发挥作用,但随着算法不确定性风险的增加与模型算法的快速迭代,这些规制工具因各种局限而疲于应对算法规制的“下半场”,产业发展呼唤算法规制范式升级。

1.个体赋权

算法权力的本质是一种现代权力范式和权力叙事。作为基于算法而产生的新型权力形态,算法权力也在呼唤新型权利保护模式。为遏制算法权力滥用所引致的负外部性,理论和实践将视野聚焦至权利进路。例如,欧盟在数据法视野下以赋予数据主体“权利束”的方式制约算法权力;美国在“权利对抗权力”的政治理念和法律架构之下,通过建构信息隐私“权利树”和“法律树”的方式,形成制衡算法权力的宪制安排。[12]个体赋权工具符合社会一般认知,秉承权利是个体应对算法等新技术风险的有力工具并对制约算法权力滥用具有工具性价值的理念。[13]但从规范效能来看,个体赋权存有如下缺陷。首先,个体赋权体现的是一种形式正义,有过于追求形式而忽略实质正义之嫌。权利是正义的表现方式,实现

正义离不开合法权利的保障。[14]但是，基于理性人假设建构的权利进路在实践中存在问题，囿于权利主体的行为偏差、乐观主义态度及高昂行权成本，个体赋权常常流于形式。其次，个体赋权的绝对化倾轧其他法益，不利于社会公共利益的维护。向理念、价值的方向思考是法益论的典型思考方向之一，[15]导致当下权利对抗权力语境下的个体赋权在绝大多数的论著中出现了绝对化的倾向，没有意识到“消费者权益保护和企业发展同为法律所欲达成的目标”。[16]最后，个体赋权中的权利合法性证成存有争议。为规训算法配置的权利绝大多数都是新兴数字权利，虽然具有社会根基，但是在实定法中缺乏明确的法律确认，在理论上难以取得共识。

2.算法透明

长期以来，算法“黑箱”被视为私人利益俘获公共利益、资本规避公权力约束的“暗室”，成为算法操纵、算法剥削等算法风险的主要根源。算法“黑箱”的本质特征在于算法自身的不透明性和难以理解性。面对密不透风的“算法围墙”，人们对于算法社会充满不确定的预期，而这种不确定预期又进一步增加了人们的恐惧感和不安全感。人类社会的文明史就是一部消除风险、增加安全的历史。为最小化风险、最大化安全，理论和实践将算法透明作为算法治理体系的重要内容。诚然，如美国大法官布兰代斯所言，“阳光是最好的消毒剂”，公开算法源代码、解释算法决策逻辑、披露算法信息这些算法透明举措对于重构公众对算法的信任、增强人们的安全感具有一定的效果。但是从治理效能来看，算法透明面临诸多困境。首先，算法透明因透明悖论而滋生规避算法、算计算法及攻击算法的风险。算法信息公开得越多，算法透明度越高，算法就越容易遭受攻击，也越容易被算法企业的对手算计，甚至被对手反向破解。其次，实践中“全有或者全无”式的算法透明与国家安全、社会秩序、私主体权利等法益相冲突。最后，算法透明忽视算法风险生成原因的动态复杂特征，未意识到强行公开算法并无益处。算法风险肇因异常复杂，如果算法风险的发生与算法建构本身无关，那么强行公开算法并无必要。

3.“以技治技”

20世纪末，美国学者劳伦斯·莱斯格在《代码》一书中提出网络规制四要素理论，其中“代码即法律”这一经典命题开创了网络法学研究之先河。[17](P6)这一理论将技术视为治理网络问题的有力工具。后续网络法学界涌现出的大量关于网络治理的研究几乎都受到了“代码即法律”这一命题的影响，着重探讨技术对于政府实现网络规制预期目标的重要性。尽管“以技治技”具有独特的激励功能，但是不加以法律归化的“以技治技”却因如下局限正走向失败。首先，不加以法律归化的技术将成为“巨吉斯之戒”，不仅对规制技术问题无益，还可能监视一切，化身刺破“技治主义”所幻想的网络自由的利器。其次，“以技治技”遵循的是自我优待而非社会共识的基本逻辑。[18]在技术创新主要由技术巨头推动的背景下，技术巨头凭借与监管部门在技术治理博弈中占据的垄断话语权优势，容易将自我利益和自我偏好植入技术治理进路，在解决问题的同时带来新的问题。最后，“以技治技”解决的问题有限，难以充分满足技术问题治理所需。该工具一般只能解决纯粹的技术问题，难以应对与价值关联密切的问题。

## 三、算法规制引入算法审计的证立

综上所述，基于回应型规制、适应型治理等理论衍生出的个体赋权、算法透明、“以技治技”等代表性算法规制工具都过分关注个别化的风险规制，缺乏整体性思维，难以充分因应算法规制需求。面对加速迭代的算法环境和频频发生的规制失灵，应对之道应是转向不仅能够彰显政府规制能力，而且能持续激发算法主体自我规制动力的基于元规制理论的算法审计。

（一）算法审计的理论指引

元规制理论滥觞于普通法系，德国将其引入社会性规制，并发展出一套完备的理论体系。不同学者对元规制的认知不一：帕克认为，元规制就是对规制者的规制过程；[19](P15)斯科特则指出，元规制注重对行为的规训和导向，是对企业内部自我规制进行的间接管制；托依布纳的观点与斯科特类似，都认为元

规制是一种间接的干预形式,并进一步表示,元规制在弥补法律规制系统中存在的局限方面具有理论优势。[20](P14)尽管学者对元规制的表述不一,但是可以看出,元规制的核心意涵在于以企业的自我规制为基础,以政府的后设规制为保障,是公权力机构针对企业自我规制开展的外部监督。在数字时代,以算法、生成式人工智能为代表的新兴数字技术的发展,重塑了信息社会技术治理的权力格局。[21]算法等新兴数字技术的广泛应用不利于规制资源的整合,致使规制资源碎片化愈发严重,而规制资源的碎片化进一步加剧了权力的分散化。相较于政府主体,市场主体凭借其在市场活动中收集的数据获得巨大的信息资源优势,并在汇聚资源的过程中产生了同政府主体掌控的正式权力相抗衡的非正式权力。相较于政府规制,企业的自我规制凭借独特的激励机制具有积极能动、快速敏捷、综合全面等优势,但囿于理性经济人追求利益最大化的行为逻辑,其亦有私利干扰、私权滥用、公责规避等难以克服的缺陷。[22]作为企业自我规制和政府规制的综合,在元规制理论指导下,政府扮演的是"掌舵者"的角色,并不对规制对象应当遵守的规制细节作出详细规定,而是要求规制对象根据行业现状与自我实际情况,制定确保内部组织合法有效运转的规制方案,既尊重市场主体自我规制的基础性地位,又利用政府的再规制消弥因市场主体自我规制缺陷而导致的市场失灵问题,为解决算法等新兴数字技术风险治理难题提供可行的理论范式和操作指南。

因此,在算法风险的规制过程中,将元规制理论嵌入算法审计之中,使算法主体在政府"注视"之下积极开展自我规制,无疑具有正当性和应然性。首先,元规制理论能够弥补前文所述的敏捷治理、场景规制等规制理论的不足。作为传统"命令—控制"型规制理论的优化产物,前文所述的敏捷治理理论、场景规制理论、回应型规制理论、适应型治理理论虽然意识到政府高权主义规制模式的不足,但仍然保持浓厚的父爱主义色彩,难以充分激励市场主体积极开展自我规制。而元规制在将监管责任"外包"给规制对象的同时又确保规制对象会受到政府监督的理念,能够在相当程度上扭转泛技术风险规制中的政府家长制作风,同时确保企业自我规制不会失序。其次,元规制理论契合算法技术的发展现状。诚如前文所述,伴随着生成式人工智能的兴起,算法技术的迭代演化周期愈发短暂,其参数体量愈发庞大,所使用的模型也愈发复杂。当作为规制问题的算法风险具有高度复杂性并处于快速变迁阶段时,外部规制者与规制对象之间存在严重的信息差,此时采取元规制监管理念更为合适。最后,元规制理念更有利于算法风险的合理分配。算法风险本质上是由内外部多种因素共同造成的,很难仅凭加强技术治理或制定硬性法律消除。算法主体是算法风险的制造者和获利者,让其承担更多的风险责任,以激励其不断完善自身算法合规体系,有助于实现风险的分配正义。

(二)算法审计正当性之确证

1.算法审计因应算法规制迭代升级的需求

尽管人类在从农业社会步入工业社会的进程中逐渐适应了风险时代下的种种不确定性,但是,2022年末以来加速演进的生成式人工智能算法,凭借超越大部分人现有认知的先进性正改变着社会对风险的认知和规制节奏。首先,区别于传统算法的参数规模,生成式人工智能算法具备千亿级别的参数,这种超大规模的参数也正是其本身的核心技术特征,例如,仅ChatGPT-3的参数规模就已经高达1750亿。就技术层面而言,算法参数规模越大,决策结果就越精准,其涌现能力和泛化能力也就越强,相应地,算法效能也就越高;但与之相伴的是,算法的可解释性下降,加重算法"黑箱"问题的严重性。有学者认为,生成式人工智能算法刷新了人类历史上的"技术黑箱"记录。[23]其次,与传统算法仅能解决特定问题相比,生成式人工智能算法具备出色的涌现能力和泛化能力,但是这种能力导致算法应用过程中产生的算法歧视、信息茧房等负外部性变得更加隐蔽而不易察知。最后,相较于传统算法的窄化应用局限,生成式人工智能算法具备的通用属性使人工智能产业链的中下游产业只需对基础算法进行微调就可以将其投入多模态跨场景的垂类应用中。这也就意味着,相较于传统算法的单一控制权,生成式人工智能算法的控制权呈分散之势,将导致生成式人工智能算法致害的责任主体变得多元且难以捕捉。

算法技术变革本身并不会成为法律规制的正当性理由，但是技术变革所引发的新风险以及新风险对现有算法规制规范、工具、力量的挑战，都呼唤着法律规制的出场，并警示人们尽快调整既有算法规制模式以充分应对技术快速变革带来的挑战。随着算法技术的迭代发展，算法规制历经沿用传统规则的算法规制 1.0 时代、确立技术中立原则和扩张平台责任的算法规制 2.0 时代，[24]以及赋予用户数字“权利束”和推动算法透明的算法规制 3.0 时代。如今，生成式人工智能算法的进一步智能化跃升需要再一次对算法规制进行迭代升级以应对算法技术的加速演进。算法审计是近年来兴起的算法规制工具，英国、加拿大、新加坡等国相继探索与算法审计相关的立法与实践活动，尝试将算法审计引入生成式人工智能时代下的算法规制实践当中，并初步取得了较好的规制实效。可以说，算法审计已经成为因应算法规制迭代升级需求的重要手段。

2.算法审计助力破解发展与安全的统筹治理困境

在《互联网信息服务算法推荐管理规定》（以下简称《管理规定》）发布之前，我国有关算法规制的规定主要集中于《中华人民共和国电子商务法》《中华人民共和国广告法》《个保法》中，对大数据杀熟、个性化推荐广告、利用个人信息进行自动化决策等侵犯公民合法权益的算法行为进行规制。伴随着算法应用过程中负外部性的日渐增大，出于对算法综合治理的需要，2022 年国家互联网信息办公室等四部门联合发布《管理规定》。《管理规定》作为我国首部以算法为专门规制对象的法律文件，围绕五类常见算法应用场景，明确了算法推荐服务提供者的信息服务规范与用户权益保护的具体要求，构建了包括算法安全风险监测与评估、算法备案管理与违法违规处理在内的多维一体的监管体系。为贯彻落实《管理规定》，加大对算法侵害公民权益行为的惩治力度，国家互联网信息办公室先后于 2022 年 4 月、2024 年 11 月开展“清朗·2022 年算法综合治理”专项行动与“清朗·网络平台算法典型问题治理”专项行动，我国算法治理发展由此逐步进入提速期。通过我国发布的算法规制文件以及开展的算法治理行动可以发现，我国在算法规制问题上采取的仍是以公权力为主的“命令—控制”型规制模式，高度依赖拥有监管职权的行政机关，通过立法或发布行政命令的方式对算法控制者施压。但是这种自上而下且由公权力主导的算法治理模式主要依靠政策推动，存在不够灵活、效率低下、治理实操性不足、不合理增加算法主体合规义务的现实问题。从根本上而言，确保算法安全几乎成为当下我国算法规制过程中的压倒性任务，算法安全的重要性在我国现有的算法规制谱系中似乎远远超过算法发展。

安全是算法时代的核心法价值，加强算法安全的出发点是正确的。但过度追求极致安全而不充分重视算法发展有失偏颇，不仅会扼杀算法技术的创新发展，损失本可实现的更高层次的个人发展和社会进步，也会因技术落后而招致更多的安全风险。据此可见，如何破解技术发展与安全的统筹治理困境已经成为数字时代亟待解决的重要问题。科技与风险相伴相生，困境的破解之法绝非简单地在安全方面增加发展砝码或者在发展方面增加安全砝码，而是需要考虑二者如何能够被一套体制逻辑统筹容纳。算法审计作为一种推动发展与安全相互促进的反身性动力机制，③不仅可以对我国现行算法规制偏重安全的现状予以纠偏，还可以很好地统筹技术发展与安全保障以实现风险的最小化和收益的最大化，更有助于回应发展与安全并重这一新时代国家治理的新要求。

3.算法审计弥合算法规制范畴内的多元利益冲突

③反身性动力机制脱胎于德国法社会学者贡塔·托依步纳的反身法理论，是在认识到以“命令—控制”模式为基础的形式法与实质法的直接控制存有局限性的前提之下，借助组织、程序、授权等框架性规制手段，注重各个主体之间去中心化的合作，在确保总体风险基本可控的前提下保持规制手段的灵活性和动态性，以此平衡法律系统和其他社会子系统之间关系的机制。本研究所构建的基于元规制理论的算法审计制度强调政府、算法企业、审计机构等多元主体间的合作，除为中风险、高风险的算法应用划定审计红线和底线外，其他风险等级的算法应用则仅提供原则性指引，在确保算法总体安全可控的前提之下为算法发展提供良好的外部监管环境，能够有效统筹推进算法技术的安全发展。有关反身法理论的研究参见：王小刚.托依布纳反身法理论述评[J].云南大学学报（法学版），2010，23(2)：107-113。

算法规制不仅需要对滥用算法的行为、对象开展必要的规制,更需要对算法规制范畴内算法利益攸关方的不同利益加以确认和平衡。在算法规制的学理和司法实践中曾出现过将算法的法律属性定性为言论、商业秘密的各种主张,以对抗针对算法开展的外部规制。例如,在美国,言论自由已经成为商业巨头对抗算法规制的一张"万能牌"。[25](P165-168)美国法官在与算法相关的众多诉讼中,选择将算法视作言论,从而基于言论自由这一宪法层面的绝对权利的不可被剥夺性驳斥每一个试图规制算法的尝试,算法结果属于言论自由的主张便在法庭的交锋中取得了胜利。与这种做法相似,有学者以算法具备秘密性、保密性、实用性、价值性特征因而符合商业秘密的构成要件为由,主张算法商业秘密保护来批判传统的算法规制路径。实践中也有公司以保护商业秘密为由对抗公开算法的诉求,并得到法院的认可。例如,在2016年发生于美国威斯康星州的涉嫌算法种族歧视的卢米斯案中,法官将涉案算法认定为商业秘密并给予保护,并声称公开算法将会严重侵犯知识产权,而知识产权的贬损又将会进一步抑制科技对人类福祉的增进作用。[25](P36)在立法层面上,随着商业秘密保护法的不断扩张,算法在不少国家的现行法秩序框架下寻得了商业秘密制度的保护。比如,德国于2019年修订的《商业秘密法》将算法纳入商业秘密的权利客体范围内。2020年,我国《最高人民法院关于审理侵犯商业秘密民事案件适用法律若干问题的规定》出台,使算法正式成为我国商业秘密制度所保护的对象。但是鉴于算法商业秘密保护与算法正义二者间的张力,把算法当作商业秘密不仅会进一步加剧算法的不可解释性,还会造成企业利益和公众权益的失衡。[26]

由此可见,之所以在算法规制的叙事脉络中会出现前述种种分歧,其根本原因在于算法规制范畴内不同利益攸关方的各种利益交错在一起,然而,既有的规制工具却往往采取非此即彼的态度,未能有效弥合算法规制下不同利益主体间的利益冲突。与既有算法规制理论指引下的规制工具相比,基于元规制理论开展的算法审计可以在对算法技术研发者的知识利益、算法产业投资者的投资利益、算法滥用侵害的用户权益以及技术发展长远视角下的社会整体利益进行承认的前提下,通过内部审计和外部审计相结合的模式,遵循保密原则,以代理审计、非侵入式审计等审计方法,不强行打开算法"黑箱"却能对算法展开细致的审查与分析,推动算法由"黑箱"阶段发展至"白箱"阶段,从而在实现对算法风险规制的同时,有效地协调算法利益攸关方的利益诉求,消弥算法规制范畴内的多方利益冲突。

4.算法审计推动事前规制与事后救济的双轨联动

法律在社会变迁所衍生的风险面前,通常采取压制性的事后救济和预防性的事前规制两种手段。前者表现为风险事故发生后,识别和确定违法行为,确定相应的责任主体和损害赔偿数额;后者则侧重于在事前采取预防性措施,以避免风险事故的发生。[27]不可否认,无论是压制性的事后救济还是预防性的事前规制,在某种程度上都能发挥相应的预期功能。比如,事后损害赔偿救济长期以来在弥补受害人损失方面发挥着举足轻重的积极作用。然而,具体到算法规制领域,针对算法侵权所展开的损害赔偿救济在落实过程中不仅面临着诉讼、举证等成本的额外支出,同时会对社会总体福祉造成较大的损耗。与算法风险损害后果的救济成本相比,针对算法采取事前的风险预防是一种前瞻性的手段,其功能优势可能会更加明显。也正因如此,预防性法律规制在全球范围内蓬勃发展,多数国家和地区积极出台预防性的法律制度并开展预防性的法律实践。例如,2024年8月1日正式生效的欧盟《人工智能法案》采取风险预防的监管思路,将人工智能可能造成的风险划分为低风险、中风险、高风险和不可接受的风险四个等级,根据风险的不同等级定制相应的事前防范措施。但是囿于风险生成因素的错综复杂和人类有限的认知能力,任何社会都不可能因规制而达到零风险状态,社会整体的监管资源是有限的,如果在事前规制方面投入过多资源,不仅会付出高昂的成本,还会间接削减事中监管和事后救济所能获得的资源。这就需要对日益勃兴的预防性事前规制加以限制,否则,各种假借预防之名所采取的措施最终将会异化为不受控制的"利维坦"。[28]

然而,事前规制与事后救济并非相互排斥、相互取代、非此即彼的关系,而是二者可以相互促进、相

互弥补、长期共存。一方面，算法审计基于分类分级原则来确定是否开展审计以及开展何种方式、何种频率的审计从而为预防性规制设上防止其异化的“枷锁”；另一方面，赋予算法审计结论以法律效力又可以使其成为监管机关对算法进行问责、算法受害群体对算法主体进行索赔的依据，从而推动事前规制与事后救济在算法规制中的双轨联动。

## 四、算法审计的规范构造

作为规制算法风险的新工具，算法审计是指在算法伦理、法律法规等算法规范指引下，利用技术性措施和非技术性措施审查输入数据、模型本身、决策影响等与算法相关的活动。算法审计具备独特的治理效能，且能弥补现有算法规制工具的不足，备受全球范围内多数国家政策制定者的关注。虽然近期多国立法对其有所言及，但是目前算法审计规则呈碎片化发展趋势，全球范围内尚未形成一套完备的、体系化的算法审计制度。结合域外立法经验以及审计学领域的审计实践，未来我国算法审计制度应围绕审计原则、审计主体、审计内容、审计方法、审计效力等几个方面展开具体的构造（图 1）。

### （一）算法审计原则

算法审计原则既是算法审计活动必须遵循的基本原则，也是解释、补充以及发展算法审计法律制度的基本原则。故而，算法审计原则对于构建算法审计制度具有重要的功能和意义。然而，既有涉及算法审计规范构造的文章鲜有提及算法审计应当遵循的基本原则。即便提及，也是以介绍传统审计学领域的知识为主，没有关注到算法审计活动本身的特殊性，这无疑是既有研究的一大漏洞。算法审计作为一种基于传统审计的新型审计方式，自然需要遵循依法审计、客观公正、保守秘密等传统审计的基本原则。但是鉴于客体的特殊性，算法审计还应遵循分类分级审计、动态持续审计、全生命周期审计、自愿审计与强制审计相结合等算法审计所特有的审计原则。首先，分类分级审计原则要求开展算法审计的方法、方式、频率应当根据算法应用所可能产生的风险等级或影响范围加以确定，针对不同类型的算法不能一概而论地开展相同的算法审计活动，否则不仅容易造成审计资源的浪费，也难以平衡算法规制范畴内利益攸关者的利益诉求。其次，鉴于生成式人工智能算法的涌现，算法技术迭代周期大大缩短，新算法新业态层出不穷，因此，算法审计应当动态持续地开展而非静态固定在算法生命周期内的某个时间节点，以便及时发现算法迭代所产生的新风险、新问题，确保审计工作的时效性。再次，一套算法模型的成功应用需要历经设计、研讨、研发、部署、投放使用、运维等多个阶段。算法全生命周期内任何一个环节的漏洞都可能引发大规模算法风险事件，这就要求算法审计的开展应当“无死角”覆盖从设计到运维的全生命周期，不遗漏任何可能的风险点，确保算法审计工作的全面性。最后，算法审计治理效能的有效发挥离不开企业自我规制的激励和政府规制的掌舵，应坚持自愿审计与强制审计有机结合的原则。一方面，有助于调动企业自主开展审计的积极性和自主性，为企业持续开展自我规制提供源源不断的激励动力；另一方面，有助于为企业自主开展的审计活动划定底线和红线，为高风险、高危害的算法应用设置“安全钳”。

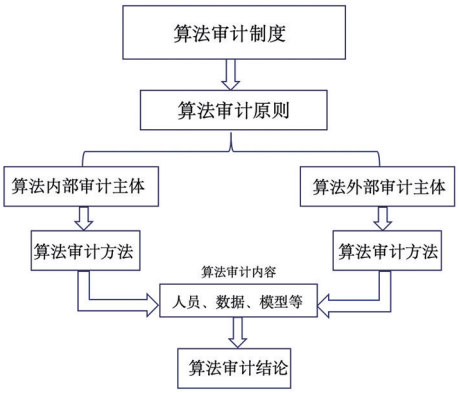


图 1　算法审计制度逻辑

### （二）算法审计主体

算法审计主体是负责组织开展算法审计活动的责任主体。根据审计活动是在算法企业内部开展还是由外部机构实施，可以将算法审计主体分为内部审计主体和外部审计主体。主体的内外划分以及二者在算法审计活动中的作用从某种意义上说也是元规制理念的缩影，因为其分别代表了算法审计中的自我规制和政府规制。算法的内部审计又被称为算法的自审计，是企业内部针对企业研发应用算法开

展的审计,由企业内部的算法审计部门成立的审计团队具体落实,在不设审计部门的情况下由审计顾问实施。例如,美国Pymetrics公司在研发招聘算法前,组织成立公司内部审计小组对招聘算法进行审计,以避免算法应用中歧视风险的发生。算法的外部审计主体主要是行政监管机关。与内部审计主体开展的审计相比,行政监管机关开展的算法审计多为事后审计,且具有个案属性。外部审计侧重于以公平性、透明度、安全性等指标评估算法内嵌社会过程的合法性、合规性,旨在为算法问责提供依据。例如,2021年,美国联邦贸易委员会发布公告称其对Everalbum公司违规搜集用户面部数据进行训练的人脸识别算法开展了审计工作。[29]此外,在美国的产业实践中,还存在类似于传统会计师事务所、律师事务所的独立运营的第三方审计组织,这类审计组织通常以接受行政监管机关、公司的委托或者出于积极承担社会责任的方式开展具体的算法审计活动。例如,美国算法正义联盟曾主动对美国市面上主流的人脸识别算法、个性化推荐算法进行审计,并以研究报告、学术论文等形式将审计结论向社会公开。[30]

(三)算法审计方法

算法审计主体在审计实践中可以根据不同算法模型选择不同审计方法,常见的算法审计方法有以下五种。第一种是代码审计。代码审计是深入检验建构算法模型的通盘代码序列的审计方法,此种审计方法不仅对审计人员的计算机综合素质提出极高的要求,还会直面算法商业秘密制度所构筑的保护屏障,对外部审计的开展构成极大的障碍。第二种是爬取审计。爬取审计是审计主体利用网络爬虫技术潜入审计对象的平台,抓取平台数据进行检测的审计方法。鉴于绝大多数商业主体为保护知识产权、数据财产权等权利对自己开发运营的平台设置了反爬技术装置,爬取审计在审计过程中会因破坏反爬技术措施而面临违法风险。第三种和第四种分别是众包审计和代理审计。这两种审计方法的原理极为相似,都是以众多算法用户的算法体验为审计样本;二者的区别在于,众包审计的审计样本由众多真人用户提供,代理审计的审计样本则是出自技术人员创设的“仿真人”。相较于代理审计,众包审计因需要招募众多真实测试用户,面临较高的操作成本;而因反“网络水军”“虚假用户”等机制的存在,代理审计面临一定的法律风险。最后一种为非侵入式审计。非侵入式审计与传统审计实践中的田野调查类似,是通过采访算法研发主体、算法用户或者直接访问用户账户与平台算法进行人机交互从而完成审计分析的一种审计方法。相较于其他几种审计方法,非侵入式审计显然会大大增加审计人员的工作量,但是非侵入式审计在实践中不仅面临的障碍较小,还能规避审计涉及的法律风险。鉴于爬取审计、代理审计涉嫌侵犯知识产权、违反网络监管规定,无论内部审计还是外部审计都应当避免在审计实践中使用这两种审计方法。

(四)算法审计内容

算法审计内容关乎算法风险诱因的具体排查,是算法审计活动的关键。算法审计在算法发展的不同阶段所关注的内容各有侧重。比如,早期的算法审计重点审查用以训练算法模型的数据的质量和模型本身的准确性等算法性能问题,而深度学习尤其是生成式人工智能算法普及以后,算法审计的关注重点就从算法性能转向模型可解释性、数据代表性等与算法效能密切相关的内容。[31]尽管随着算法技术的迭代更新,数据和算法模型间的关系变得愈发复杂,算法风险肇因也变得扑朔迷离,但总体而言,算法审计内容主要还是围绕人员、数据、模型以及与之相关的活动展开。以审查算法有无歧视风险为例,首先,算法工程师出于维持积极的自我认同以及维护自尊的需要,[32]通常会不自觉地将内隐偏见通过代码形式外化,从而产生算法歧视。在此场景下,对算法工程师的调查就显得尤为重要。其次,训练算法模型的数据代表性不足、过足以及数据本身被污染、被“投毒”,都会产生数据维度的算法歧视。因此,审查模型算法训练数据集的代表性和完整性就成为审计活动的关注重点。再次,算法模型中目标函数的不准确将会直接影响算法能够找到的“解”的质量。换言之,如果算法模型采取了有价值偏向的目标函数,那么算法结果就会朝着追逐“不当利益”的方向发展。所以,算法模型本身也是审计人员不可回避的

审计对象。最后，在算法模型训练以及运维过程中会出现属性噪声、标签噪声等各种噪声，这些噪声会使算法输出的结果偏离预设的目标。所以，与算法研发、部署、应用等相关的活动人员、活动内容也应被纳入算法审计的内容范围。

（五）算法审计效力

为充分发挥算法审计在衔接算法风险事前规制与算法侵害事后救济双轨联动中的抓手作用，赋予算法审计以法律效力就成为整个算法审计规范构造中不可或缺的一环。首先，当算法审计主体检测出算法存在导致或可能导致大数据杀熟、信息茧房的风险时，有权要求算法服务提供者对此作出解释并为其提供的算法服务补充修改完善建议。如果审查发现算法隐含重大风险隐患，审计主体有权将审计报告上报给当地网信、公安、市场监管等责任主管部门。其次，行政监管机关可以根据其自行组织或者委托第三方审计组织开展的审计活动出具的审计报告对违法违规的责任主体进行问责。比如，当审计报告显示算法服务提供者在训练算法模型过程中存有违反个人信息保护义务和数据安全保护义务的违法违规情形时，行政监管机关可以根据《个保法》《中华人民共和国数据安全法》中的个人信息处理和数据处理的行政责任条款对算法服务提供者实施具体的行政处罚。最后，为在具象微观层面上对遭受算法侵害的个体提供个案正义意义上的预防和救济，[33]算法审计报告可以成为被侵权人主张侵权损害赔偿的证据，但是侵权损害赔偿责任成立与否还需要看是否满足侵权责任的法定构成要件。

综上所述，结合算法审计的制度逻辑，在元规制理念的指引下，未来我国算法审计实践可考虑预先由公权力主体制定算法应用风险等级清单，④为中风险、高风险的算法应用划定审计红线和底线，为其他风险等级的算法应用提供原则性指引。建立算法审计“白名单”制度并将其融入社会信用体系建设，鼓励企业在操作指引下根据行业发展现状与自我定位制定企业内部审计方案，积极主动开展相应的算法审计活动（表 1）。

**表 1　算法审计模式**

| 风险等级 | 内部审计 | 外部审计 | 强制与否 | 是否对外发布审计报告 |
|---|---|---|---|---|
| 最小风险 | 无须审计（或自愿审计） | 无须审计 | 不强制 | 不作强制要求 |
| 低风险 | 定期审计 | 无须审计 | 不强制 | 不作强制要求 |
| 中风险 | 定期审计 | 发生安全事故等情况时进行 | 强制 | 不作强制要求 |
| 高风险 | 持续审计 | 持续审计 | 强制 | 是 |

## 五、余论：提升算法审计的可操作性

数字时代的号角已然吹响，算法引擎在号角声中正加速推进人类社会从以土地为资源的农业社会、以矿产为资源的工业社会向以数据为资源的算法社会变迁。当下，算法与数据已经成为推动新质生产力形成的关键构成要件。从另一角度来看，这也可能演化为算法异化风险之基。透视算法运行逻辑、认清算法技术本质、探寻算法治理之道将会是数字时代法学面临的一项长期性、前沿性的时代命题。

算法审计作为算法规制的一项创新工具，其不仅可以有效弥补现有算法规制工具的不足，还有助于推动法律、规范、技术等多元共治的实现，因而逐渐获得学界和实务界的认可。但是算法审计作为新兴事物，还面临审计指引空缺、审计成本较高、审计人员匮乏等可操作性难题。能否克服算法审计面临的

④关于算法应用风险等级划分，目前理论和实践尚无定论，可考虑借鉴欧盟《人工智能法案》关于人工智能风险等级划分的经验，结合我国《数据安全技术 数据分类分级规则》《金融数据安全分级指南》等规范性文件有关数据分类分级的方案，以算法应用产生的风险可能对公民个体权利、社会公共利益及国家安全造成的影响为标准，将诸如垃圾邮件过滤算法、电子游戏调配算法等算法应用可能产生的风险划分为最小风险；诸如自动回复机器人、聊天机器人等算法应用可能产生的风险划分为低风险；诸如招聘简历筛选算法、人脸识别算法等算法应用可能产生的风险划分为中风险；诸如社会信用评估算法、司法裁判辅助算法等算法应用可能产生的风险划分为高风险。

可操作性难题将深刻影响算法审计治理效能的释放以及算法审计在算法规制谱系中所占据的地位。基于此,未来我国在构建算法审计制度时需要考虑加速算法审计的规则供给,加快完善算法审计的法律体系,及时明确算法审计的法律地位,对算法审计人员的从业资格、执业规范以及责任义务作出细化的规定。探索建立类似社会信用体系建设"白名单"制度的算法审计"白名单"激励性机制,以充分激励企业算法治理合规,加快算法审计人才队伍建设,充实算法审计人才储备,以便加快探索出与算法技术发展趋势相适应且具备实操性、行之有效的审计路径,积极引导算法服务向着"以人为本,智能向善"方向发展,以算法审计促进算法产业的健康发展。但是需要指出的是,在一套行之有效的算法审计制度落地之前,传统算法规制理论与规制工具将会继续发挥规制作用。此外,即便算法审计真正落地,也不能保证其没有缺陷,应当融合好各种规制理论指导下的规制工具,充分发挥各种规制范式的优势。同时,元规制理论指导下的算法审计也应根据审计实践适时调整相应的指引策略。唯有如此,方能探索出一条法律与科技良性互动的康庄大道。[34]

**参考文献:**


[1] 郑智航. 当代中国数字法学的自主性构建[J]. 法律科学(西北政法大学学报),2024,24(2):81-93.

[2] 张凌寒,于琳. 从传统治理到敏捷治理:生成式人工智能的治理范式革新[J]. 电子政务,2023,(9):2-13.

[3] Goldsmith J,Wu T. Who controls the internet? Illusions of a borderless world[M]. Oxford:Oxford University Press,2006.

[4] 李晟. 国家安全视角下社交机器人的法律规制[J]. 中外法学,2022,34(2):425-444.

[5] World Economic Forum. Agile governance:reimagining policy-making in the fourth industrial revolution[EB]. https://www3.weforum.org/docs/WEF_Agile%20Governance_for_Creative_Economy_4.0_Report.pdf,2019-10-01/2024-09-28.

[6] 赵精武. "元宇宙"安全风险的法律规制路径:从假想式规制到过程风险预防[J]. 上海大学学报(社会科学版),2022,39(5):103-115.

[7] 赵静,薛澜,吴冠生. 敏捷思维引领城市治理转型:对多城市治理实践的分析[J]. 中国行政管理,2021,(8):49-54.

[8] 贾开,赵静,傅宏宇. 应对不确定性挑战:算法敏捷治理的理论界定[J]. 图书情报知识,2023,40(1):35-44.

[9] 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,2024,36(1):261-279.

[10] 郭雳. 精巧规制理论及其在数据要素治理中的应用[J]. 行政法学研究,2023,(5):26-39.

[11] 张涛. 人工智能治理中"基于风险的方法":理论、实践与反思[J]. 华中科技大学学报(社会科学版),2024,38(2):66-77.

[12] 余成峰. 信息隐私权的宪法时刻规范基础与体系重构[J]. 中外法学,2021,33(1):32-56.

[13] 崔聪聪. 数据限制处理权的法理基础与制度建构[J]. 比较法研究,2022,(5):75-88.

[14] 谷佳慧. 数字时代正义的内涵变迁及法治保障[J]. 北方法学,2023,17(5):131-145.

[15] 高志宏. 大数据时代个人信息保护的理论反思与规则优化[J]. 学术界,2023,(7):122-137.

[16] 林洹民.《个人信息保护法》中的算法解释权:兼顾公私场景的区分规范策略[J]. 法治研究,2022,(5):48-58.

[17] 劳伦斯·莱斯格. 代码 2.0:塑造网络空间的法律(修订版)[M]. 李旭,沈伟伟,译. 北京:清华大学出版社,2018.

[18] 郑智航. 网络社会法律治理与技术治理的二元共治[J]. 中国法学,2018,(2):108-130.

[19] Parker C. The open corporation: effective self-regulation and democracy[M]. Cambridge: Cambridge University Press, 2022.

[20] 科林·斯科特. 规制与法律:前沿问题研究[M]. 安永康,译. 宋华琳,校. 北京:清华大学出版社,2018.

[21] 谢永江,杨永兴. ChatGPT 法律风险及其规制[J]. 南京邮电大学学报(社会科学版),2023,25(5):29-39.

[22] 黄文艺,孙喆玥. 论互联网平台治理的元规制进路[J]. 法学评论,2024,42(4):111-122.

[23] 王洋,闫海. 生成式人工智能的风险迭代与规制革新——以 ChatGPT 为例[J]. 理论月刊,2023,(6):14-24.

[24] 张凌寒. 算法规制的迭代与革新[J]. 法学论坛,2019,34(2):16-26.

[25] Pasquale F. The black box society: the secret algorithms that control money and information[M]. Cambridge: Harvard University, 2016.

[26] 许中缘,郑煌杰. 生成式人工智能算法专利保护的理据与进路[J]. 贵州师范大学学报(社会科学版),2024,(1):135-147.

[27] 朱荣荣. 类 ChatGPT 生成式人工智能对个人信息保护的挑战及应对[J/OL]. 重庆大学学报(社会科学版),1-14[2025-09-05]. https://link.cnki.net/urlid/50.1023.c.20230921.1151.002.

[28] 黄文艺. 论预防型法治[J]. 法学研究,2024,46(2):20-38.

[29] 张欣,宋雨鑫. 算法审计的制度逻辑和本土化构建[J]. 郑州大学学报(哲学社会科学版),2022,55(6):33-42.

[30] 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.

[31] 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.
[32] 李成. 人工智能歧视的法律治理[J]. 中国法学，2021，(2)：127-147.
[33] 王莹. 算法侵害责任框架刍议[J]. 中国法学，2022，(3)：165-184.
[34] 杨永兴. DeepSeek 等开源模型法律风险治理研究[J]. 四川轻化工大学学报(社会科学版)，2025，41(4)：27-38.

# The Path Shift of Algorithm Regulation: Algorithm Audit Based on Meta Regulation Theory

Yang Yongxing
(School of Law, Ocean University of China, Qingdao 266100, China)

**Abstract**: The accelerated evolution of algorithm technology calls for iterative upgrading of algorithm regulation. The traditional algorithmic regulatory tools derived from agile governance theory and scenario regulation theory, such as individual empowerment, algorithm transparency, and technology governance, not only cause conflicts between regulatory subjects and regulatory recipients, but also struggle to cope with the regulatory challenges brought by generative artificial intelligence algorithms. The algorithm audit based on meta regulatory theory, as a reflexive driving mechanism, can effectively compensate for the shortcomings of traditional algorithm regulatory tools, eliminate conflicts of interest among stakeholders within the scope of algorithm regulation, and help enhance the international influence of China′s algorithm governance paradigm. Therefore, algorithm auditing guided by the theory of meta regulation should become the necessary shift in China′s algorithm regulation path in the future. Based on the reference to foreign legislative experience and combined with traditional auditing practices in the field of auditing, the standardized construction of algorithmic auditing in China in the future should include auditing principles, auditing subjects, auditing methods, auditing content, auditing effectiveness, and other aspects of algorithmic auditing.

**Key words**: algorithm; algorithm regulation; algorithm auditing; algorithm specifications

责任编辑：王明舜