中州学刊

论新闻生产算法影响评估机制的构建

张　超　　陈　莎

摘　要：智媒时代算法已广泛嵌入新闻生产全流程，与之伴随的各类风险值得关注。当前国家层面的算法风险治理侧重具有舆论属性或社会动员能力的推荐算法和深度合成算法，并未将新闻生产相关的其他算法纳入其中，形成算法风险治理的盲区。新闻生产与公共利益密切相关，构建专门的面向新闻生产的算法影响评估机制势在必行。算法影响评估具有治理对象细分、治理节点前移、评估要素全面等优势。对新闻生产而言，算法影响评估可以提升新闻生产算法风险治理的针对性、公平分配风险责任、有效保护新闻用户权益、提高新闻从业者算法素养。新闻生产算法影响评估应根据算法风险影响对象识别风险类别、综合多重因素确定风险等级、基于风险等级确定责任义务来构建具体的评估机制。

关键词：算法影响评估；算法风险；风险为本；算法素养

中图分类号：G206.2　　**文献标识码：**A　　**文章编号：**1003-0751(2024)02-0168-09

智媒时代作为扮演媒介基础设施角色的算法已广泛嵌入新闻生产全流程。算法在提升新闻生产效率的同时也带来失实风险、决策风险、侵权风险等各种算法风险(algorithmic risks)。例如，美国《洛杉矶时报》误报当地发生地震，路透社新闻线索发现系统 News Tracer 因新闻事件未达到事先设定的数据量而未将其判定为新闻选题，BuzzFeed 用随机森林算法调查美国空军秘密活动将跳伞识别为飞机，一些自动化新闻写作算法存在“洗稿”和文本质量不高等问题。针对近年来兴起的生成式人工智能(Generative AI)热潮，英国《卫报》明确表示该报将防范生成式人工智能工具及其底层训练集中蕴含的偏见危险[1]。科技杂志《连线》(*Wired*)表示不会发表由人工智能生成的报道[2]。

算法风险是一种高技术风险，具有隐蔽性、系统性、批量性和不可避免等特点。当前国内新闻传播领域的算法风险治理研究较多关注个性化推荐算法，讨论“信息茧房”“算法歧视”等风险，在治理上多探讨算法伦理问题。在实践中，国家层面较关注具有舆论属性或社会动员能力的推荐算法和深度合成算法。2022 年 3 月 1 日施行的《互联网信息服务算法推荐管理规定》要求：具有舆论属性或者社会动员能力的算法推荐服务提供者应当按照国家有关规定开展安全评估[3]。2023 年 1 月 10 日施行的《互联网信息服务深度合成管理规定》要求：具有舆论属性或者社会动员能力的深度合成服务提供者，应当按照《互联网信息服务算法推荐管理规定》履行备案和变更、注销备案手续[4]。新闻生产算法中除了推荐算法和深度合成算法，在采、编、播、审、发各环节中都有算法应用，但这些算法没有纳入算法风险治理体系，形成新闻生产算法风险治理的盲区。

现阶段，只有极个别大型新闻媒体具备一定的研发能力，更多的新闻媒体要么将算法研发外包于技术公司，要么直接引进使用，驾驭算法的技术能力

收稿日期：2023-08-20

基金项目：国家社会科学基金项目“智媒时代新闻生产算法风险及其协同治理研究”(19BXW020)。

作者简介：张超，男，山东大学文化传播学院教授、博士生导师(山东威海　264209)。陈莎，女，山东大学文化传播学院新闻传播研究所研究助理、博士生(山东威海　264209)。

不足。如何有效防范新闻生产算法风险,构建专门的、面向新闻生产的算法风险评估与治理机制势在必行,算法影响评估(algorithm impact assessment,简称 AIA)值得借鉴。

在新闻生产领域构建算法影响评估机制并不是简单地效仿,需要对算法影响评估进行场景化改造。本文首先探讨算法影响评估在算法风险治理中的优势,分析构建新闻生产算法影响评估机制的现实意义,最后结合中国语境和"风险为本"理念提出具体的新闻生产算法影响评估机制,为新闻生产算法风险治理提供新思路。

一、算法影响评估的内涵

算法影响评估是特定主体根据已确定的标准对算法自动化决策系统的设计、应用和数据处理等内容进行全面评估,据此确定该系统对特定领域内的个人或群体所产生的影响程度和风险等级,以寻求减缓、消除负面影响和风险应对方案(措施)的算法治理活动[5]。算法影响评估最初面向政府使用的公共服务算法,主要用于评估算法风险,维护公共利益,保障社会公平和公正。现在,算法影响评估的对象已从公共服务算法拓展至涉及公共利益的部分商业平台算法。新闻生产涉及公共利益且影响广泛,算法影响评估自然也可以应用于新闻生产的算法风险治理。

算法影响评估蕴含"风险为本"(risk-based approach,又译为风险导向、风险为基)的理念。"风险为本"是一套将风险分析逻辑放置在风险构成要素与风险发生场景之中的风险治理框架,目标不是消除风险,而是将风险作为治理对象,管理并降低风险[6]。它摈弃传统路径对风险全有或全无的"二元化"判断,转而进行"程度性"评估,以个案分析的精神在相应场景中评估风险大小[7]。当识别出风险等级后,治理主体按照"风险为本"的路径会分配较多资源处理高风险议题,将次要资源分配给中风险议题,对低风险议题则采取容忍,或是分配最少资源处理[8]。

算法影响评估已在多个国家发展为专门的算法影响评估制度。2019 年,加拿大率先在《自动化决策指令》(*Directive on Automated Decision-Making*)中提出算法影响评估框架和方案;2022 年,欧盟、美国分别在《人工智能法案》(*Artificial Intelligence Act*)、《算法问责法案(2022)》(*Algorithmic Accountability Act of* 2022)中阐述了算法影响评估细则。在算法技术迭代发展、算法权力愈发膨胀的背景下,算法影响评估显得愈发重要和必要。

二、算法影响评估在算法风险治理中的优势

算法影响评估之所以被纳入一些国家和地区的算法风险治理体系,独特的治理优势是主要原因。

1.治理对象细分:针对特定算法决策系统

算法风险治理需要考虑治理成本。当算法成为社会运行的基础设施时,政府对所有算法进行监管既不现实,也无必要。算法影响评估针对的是特定应用场景下被评定为特定等级风险的算法。

在欧盟,特定算法决策系统是指高风险算法。欧盟《人工智能法案》根据对个人基本权利威胁程度的大小将算法分为禁止、高风险、有限风险、最小风险四个风险等级。其中,高风险算法指在重要基础设施(如交通)、教育或职业培训、可能干涉人的基本权利的执法(如评判证据可靠性的系统)等场景中的应用系统。美国《算法问责法案(2022)》针对的是"增强的关键决策过程"(augmented critical decision processes)。"增强的关键决策过程"是指一种采用自动化决策系统进行关键决策的程序、过程或其他自动化决策活动,其中的"关键决策"是指对消费者的生活产生任何法律上的、实质性的或类似重大影响的决定或判断。该法案规定部署"增强的关键决策过程"的特定实体要开展算法影响评估,包括评估任何已知的危害、缺点、故障案例,对自动决策系统或增强型关键决策过程的隐私风险和隐私增强措施进行持续测试和评估[9]。

从监管层面看,由于只对特定算法进行影响评估,与公共利益密切的各类算法是被监管的重要对象,这提醒相关主体要对这类算法重点关注。由于不要求所有算法参与评估,既给了非特定风险算法相对宽松的创新发展空间,也大大降低了社会治理成本。算法影响评估由于要求特定主体定期向算法监管机构、社会披露特定算法的风险,可以克服算法监管中的信息不对称,使作为外部监管者的算法监管机构得以洞悉算法黑箱进而全面了解算法的运行状况和风险,既补强了算法监管机构的监管能力,也缓解了监管压力[5]。

2.治理节点前移:由事后追责转向事前评估

从技术的路径依赖(path-dependency)来看,一

旦新兴技术被大规模推广和采用,便会产生锁定(lock-in)效果,当技术在市场渗透后,政府才可能因负面外部性问题关注技术风险,此时的管制将面临极大阻力[10]。传统的技术风险治理多采取事后追责,待风险发生并造成实质损害后,相关救济、治理措施才会启动,事后追责暗含的逻辑是“不出事等于没风险”。

算法风险是客观存在的,也是无法避免的,算法影响评估不是消除风险,而是在可预见的范围内采取适当的措施将风险降至最低,重在预防。算法影响评估对风险的预防主要体现在两方面:一是通过算法影响评估了解算法存在的风险;二是评估针对风险所采取的缓解措施,让风险可控。风险评估均需在算法部署前落实,形成算法影响评估报告,评估报告既要向公众披露,也要向监管机构报备。披露与报备反过来又会倒逼相关主体在算法设计、开发过程中采取更为严格、全面的风险防范措施。

3.评估方式动态:对算法影响周期性评估

算法是动态、迭代的,并非一成不变的。算法影响评估考虑到算法在实际应用中的动态性特征,提供了周期性评估的思路。加拿大《自动化决策指令》规定,在生产任何自动化决策系统之前要完成算法影响评估,当系统功能或自动化系统的范围发生变化时要更新算法影响评估[5]。欧盟《人工智能法案》规定,高风险人工智能系统提供者不仅要在设计、研发阶段即采取措施确保系统可识别未来的风险,还要在人工智能系统上市后持续进行风险识别与判断[11]。已经接受过合格评估程序的高风险人工智能系统进行实质性修改时,也需要重新评估。

三、构建新闻生产算法影响评估机制的现实意义

笔者在调研中发现,一些新闻从业者对算法风险的认知存在两种极端:一种是盲目信任算法,不质疑各类风险问题;一种是不信任算法,对算法的接纳非常谨慎,甚至消极。如果对算法风险视而不见,损害的是新闻业的公信力;如果对算法技术因噎废食,新闻业的发展将错失技术红利,停滞不前。智媒发展势不可挡,新闻业需要直面人机协作的现实,善用算法,让算法给新闻业的价值最大化,算法影响评估机制的构建有着现实意义。

1.提升新闻生产算法风险治理的针对性

从目前全球的算法风险治理实践看,美国没有专门针对新闻生产算法风险的治理措施。欧盟虽然推出了算法影响评估,但与新闻生产相关的算法属于何种风险等级还不可知。有研究指出,欧盟的政策文件倾向于将讨论对象停留在抽象层面,不提及媒体和新闻业,但内容往往与之有关[12]。在中国,具有舆论属性或社会动员能力的推荐算法和深度合成算法已被纳入监管范围,但新闻生产的其他算法还没有进入监管视野,从引进、部署到维护,身处技术下游的新闻媒体对算法权力及其蕴含风险的了解并不充分。或许有人会认为,新闻生产算法的本质是算法,统辖在国家层面的算法风险治理体系中即可,但这忽视了算法应用的场景性,同一种算法,应用场景不同,风险等级可能不同。例如,欧洲广播联盟针对欧盟《人工智能法案》提出异议:生成复杂文本以及生成、操纵图像、音频或视频内容的人工智能系统不应被默认归类为高风险,因为就公共服务媒体而言,这些系统经过编辑的严格审查,编辑也对产生的结果承担责任[13]。由于新闻生产场景的特殊性,新闻生产算法不能和其他领域算法混在一起无差别对待。

对公共利益的承诺是新闻业的立身之本。提供有新闻价值、客观准确的报道是新闻业承担社会责任的体现,新闻生产中一系列把关程序都是因此而设,当新闻生产权力全部或部分让渡给算法,算法能否很好地担当此任,让人心存疑虑。以技术逻辑“重新定义”的新闻实践并不一定将新闻公共性考虑在内,二者可能产生冲突。笔者在调研中发现,从算法开发者的视角来看,新闻生产算法和其他行业算法鲜有差异,算法研发会最大程度地考虑研发成本和适用领域,较少考虑新闻媒体的特殊因素,即便考虑,新闻媒体付出的成本也将是巨大的。“技术行业不会有人去为传统新闻媒体量身打造一个东西,传统新闻媒体付不起这个钱。”① 这就导致新闻生产算法很大程度上是由新闻业的“局外人”所形塑。此前,BBC和其他公共服务媒体依靠外部商业承包商来建立推荐系统。然而,这意味着这些媒体几乎无法控制所使用的数据和算法[14]。在国内,面对平台推荐算法“流量至上”的价值观,业界和学界提出研发“党媒算法”“主流算法”,将主流媒体的价值观嵌入算法,如果没有相应的评估机制,何以保证嵌入算法的价值观是相匹配的?

新闻生产环节不同,算法风险亦有差异。用于新闻媒体内部的新闻线索发现算法即便出问题也属于媒体内部问题,虽然系统可能误报,但在层层把关

之下,线索误报风险会在新闻媒体内部消除。自动化新闻写作直接生成内容,一旦出错,将是系统性、大面积地传播假新闻,由于直接面向用户,这类风险较之新闻线索误报会带来更大的负面影响,风险等级更高。

算法影响评估由于将治理对象细分,可以从两个层面提升算法风险治理的针对性:一是将新闻生产算法与其他领域算法区分开;二是将不同新闻生产环节的算法区分开。

2.公平分配风险责任

在算法构造的风险体系中,利益相关者包括算法开发者、算法服务提供者、算法服务使用者(如新闻媒体、新闻从业者)、算法服务影响者(如新闻媒体、新闻从业者、用户)、监管者等。算法风险如何分配和承担,出现问题后谁应该负责?这些都是现实问题。德国记者协会强调,新闻机构对其内容负责,编辑部门应该为涉及人工智能的新闻内容建立规范的接受和批准程序[15]。但现实问题是,很多新闻媒体没有评估技术风险的能力,如果是这样,这些媒体是否就不能使用人工智能?新闻媒体应该对内容负责,但是否只有新闻媒体要对此负责?如果推给算法本身显然不负责任,如果只归咎于新闻媒体、算法开发者或算法服务提供者同样有失公允,这样会导致新闻媒体不敢使用、算法开发者不敢开发、算法服务提供者不敢提供等问题。

解决以上问题需要公平分配风险责任,如何公平分配?需要采取对称性分配原则:第一,风险分配要与生产风险的责任相对称;第二,风险分配要与获取风险的收益相对称;第三,风险分配要与抵御风险的能力相对称。要避免让引发风险的责任主体逃脱风险的分配,或者将过少的风险分配给引发风险的责任主体[16]。算法影响评估可以披露一系列与风险、风险管控有关的重要细节,在一定程度上解决利益相关者对算法风险的信息不对称问题,有助于落实风险的对称性分配问题,也使责任认定相对明确,倒逼相关主体负责任地开发和创新,规避“有组织地不负责任”发生。

3.有效保护用户权益

技术的形成和特定的利益相一致,处于技术网络中的人都有一定的利益诉求,但利益不可能针对整个社会群体,技术在满足一部分群体利益的同时,必然有另一部分人的利益被忽视或被压抑[17]。对公民权利、自由和隐私的威胁是美国规制人工智能时最重要的考虑因素[18]。这些问题也日益受到全球重视。如今新闻媒体不仅生产内容,还要处理数据,一些大型新闻媒体甚至会成为数据中心,自然会涉及用户隐私权、算法知情权等多项用户权益保护问题,如何保护,光靠自律是远远不够的。

首先,新闻生产算法可能因过度收集用户个人信息有侵犯个人隐私的风险。新闻推荐、用户分析等需要大量用户个人数据,新闻媒体可能出于商业目的过度采集,由于算法采集用户个人信息的行为不可见,即便用户怀疑、察觉到,也可能无力维护自身权益。

其次,用户在接触各类新闻生产算法时,既有知晓算法存在、算法风险的知情权,也有不受算法决策影响的选择权,这类权益保障在国内外算法规制中均有专门说明。然而,媒体出于私利考虑,可能会有意不披露,或用技术手段阻挠用户的算法知情权。前者如美国科技媒体 CNET 发布数十篇由人工智能生成的报道,却用人类作者身份“假冒”[19];后者如算法工程师所言:“平台会将关闭算法、个性化推送的设置按钮埋藏很深,需要用户费力去找到这个设置,要多点击几次才能关闭。”②

现代技术的主要缺陷是受其影响的人对技术的设计和运行的控制权很少甚至没有,改变这种状况需要更多的人对技术的规划、设计和运行有更多的知情权和控制权,而不能仅局限于专家[20]。算法影响评估一方面评估新闻用户的算法权益是否受到损害,另一方面让缺乏相应知识的新闻用户了解算法自身存在的风险,增强用户的算法意识(algorithmic awareness),可以最大化维护用户的权益。

4.提高新闻从业者算法素养

算法素养(algorithmic literacy)是了解算法在在线应用程序、平台和服务中的使用,了解算法如何工作,能够批判性地评估算法决策,以及拥有处理甚至影响算法操作的技能[21]。当前全球范围内新闻从业者的算法素养都是比较低的。研究显示,BBC 记者对人工智能及其与新闻业关系的认识和理解有限,在讨论人工智能和算法时使用猜测和想象,并渴望了解更多[22]。

算法影响评估要求算法服务提供者提供有关算法风险的评估报告,使用该服务的新闻媒体要对使用的算法有足够的了解,了解算法系统的工作原理,认识到潜在的偏见、缺陷和风险,并将这些信息传达给新闻从业者,让其批判地使用算法系统,提升应对算法风险的能力,并为未来参与算法迭代设计创造前提,客观上提升新闻从业者的算法素养。

四、新闻生产算法影响评估机制的构建

构建新闻生产算法影响评估的起点是确定要评估的算法。从研发角度而言,算法可分为基础算法和场景算法。基础算法是为解决一类问题而研发的算法,应用场景广泛。在此基础上,算法开发者会根据客户具体业务需求研发场景算法。新闻生产算法影响评估针对的是场景算法。业务场景只是新闻生产算法影响评估的一个考虑因素,新闻生产算法影响评估机制还要考虑实体等诸多因素。笔者根据中国语境和"风险为本"理念尝试从风险类别、风险等级、责任义务三个维度构建场景化、精准化的新闻生产算法影响评估机制。

1.根据算法风险影响对象识别风险类别

《关于加强互联网信息服务算法综合治理的指导意见》指出,推进算法分级分类安全管理,有效识别高风险类算法,实施精准治理[23]。对新闻生产算法风险分类分级是实现精准监管的前提,但如何分级分类,现有政策法规并未给出具体标准;什么属于"高风险类算法",也没有得到明确界定。国外算法风险的分类多是一种概括式分类,并将分类等同于分级,如欧盟人工智能风险等级框架以危害严重性来分类,实际上是分级。

如何对算法风险进行分类?如果按照新闻业务场景对各个环节的风险进行分类,新闻生产风险有失实风险、决策风险、偏见风险、固化风险、隐私风险、媒体声誉风险、权力让渡风险、质量风险、侵权风险等,但这种分类方式比较具体。随着算法技术的发展,新闻业务场景内部会更加细分和动态变化,不利于分类的稳定性。算法影响评估中的"影响"指向新闻生产系统中受影响的利益相关者,如果根据算法风险影响对象识别风险类别,则可以解决以上问题。

依据算法风险影响对象,风险类别可分为内部风险和外部风险。内部风险是指影响在新闻媒体内部的风险,算法影响对象是新闻媒体和新闻从业者,这类风险发生在新闻生产前端,由于编辑部后续有人工把关环节,影响范围大多会控制在新闻媒体内部,对外界而言,往往不可见和不可知。外部风险是指影响在新闻媒体外部的风险,算法影响对象是个体和社会,通常靠媒体内部把关难以避免,这类风险发生时就意味着风险已经"溢出"编辑部。内部风险和外部风险的子类在不同业务场景中可能涉及前文所述九种风险的一种或几种。不同场景的影响对象、风险类别和具体的风险子类见表1。

表1 新闻生产算法应用场景与风险

生产环节	应用实例	算法技术	风险类别	影响对象
新闻采集	新闻线索发现算法	通过机器学习、数据挖掘算法,监测数据异常,寻找变化,提供决策判断	内部风险(决策风险)	新闻媒体、新闻从业者
新闻制作	自动化新闻写作算法	依赖自然语言处理算法、数据采集算法,对数据进行收集、清洗、结构化处理与分析,套用模板生成新闻	外部风险(失实风险、质量风险、媒体声誉风险)	用户、社会
	深度合成算法	利用深度合成技术实现图像、声音、文字的合成	内部风险(侵权风险)	新闻媒体、新闻从业者
新闻分发	新闻推荐算法	基于用户、内容、分发算法三要素,将内容与用户匹配	外部风险(偏见风险、固化风险、隐私风险、媒体声誉风险)	用户、社会
事实核查	自动化事实核查算法	基于自然语言处理算法、社会情境、区块链技术等,对数据进行溯源、比对、核实、评判、校验	内部风险(失实风险)	新闻媒体、新闻从业者

当然,风险类别的区分是相对的,内部风险不可控就会变成外部风险。针对不同类别的风险,评估重点也略有差异。针对内部风险,需要视风险等级重点评估与新闻媒体和新闻从业者相关的具体风险问题和缓解措施;针对外部风险,需要视风险等级重点评估与公众相关的具体风险问题和缓解措施。

2.综合多重因素确定风险等级

风险等级的确定因素是复杂的,高、中、低三个风险等级的边界应当是较为明晰的,不仅涉及新闻业务场景、影响对象,还涉及实体因素,即算法服务提供者和使用者的规模和影响力。

实体可分为新闻媒体和媒体算法服务提供者。需要指出的是,具备算法开发和服务能力的新闻媒体,既是算法服务提供者,也是算法服务使用者;不具备算法开发和服务能力的新闻媒体,属于算法服务使用者。媒体算法服务提供者是指除此之外的、

为媒体提供算法服务的实体，如新华智云、百度智能云等，它们批量研发、推广算法。

实体的规模和影响力不同，对算法风险等级的判定也可能不同。在影响力的量化上，新闻媒体综合平台年度日活用户数、媒体级别指标，可分为一类、二类、三类。一类主要是年度平均日活用户超千万的新闻媒体；二类为年度平均日活用户超百万或省级及以上级别的新闻媒体；三类为年度平均日活用户低于百万、地市级及以下的新闻媒体。

媒体算法服务提供者可根据其市场规模做二级分类。市场规模主要参考智能媒体市场年度占有率和新闻媒体服务占有率两个指标。根据国际数据公司（IDC）发布的《2021 中国智能媒体方案市场分析》[24]和中国记协发布的《中国新闻事业发展报告》[25]，可将市场规模占比 10%或服务的新闻媒体用户占比 10%以上的媒体算法服务提供者定位为大规模媒体算法服务提供者，如百度智能云（市场占比 18.1%）、新华智云（服务新闻媒体用户占比超 30%）；市场规模占比均低于 10%的则是小规模媒体算法服务提供者（见表 2）。根据智能媒体市场的发展概况，这一百分比的约定可以视实际状况进行动态调整。

表 2　新闻生产算法影响评估中的实体分类

实体	类别	判定标准
新闻媒体	一类新闻媒体	年日活用户超千万
	二类新闻媒体	年日活用户超百万或省级及以上级别
	三类新闻媒体	年日活用户低于百万或地市级及以下
媒体算法服务提供者	大规模媒体算法服务提供者	市场规模或服务的新闻媒体用户占比大于等于 10%
	小规模媒体算法服务提供者	市场规模、服务的新闻媒体用户占比均低于 10%

说明：年日活用户数均为上年度在中国的用户数。

综合新闻业务场景、影响对象、算法服务提供者、算法服务使用者和算法风险类别等因素，新闻生产算法风险可分为高风险、中风险、低风险三级（见表 3）。

表 3　新闻生产算法影响评估中的风险等级

新闻生产算法	算法服务提供者/使用者	风险等级
新闻推荐算法	一类新闻媒体；大规模媒体算法服务提供者	高
自动化新闻写作算法	一类新闻媒体；大规模媒体算法服务提供者	
深度合成算法	所有实体	
新闻推荐算法	二类新闻媒体；小规模媒体算法服务提供者	中
自动化新闻写作算法	二类新闻媒体；小规模媒体算法服务提供者	
新闻推荐算法	三类新闻媒体	低
新闻线索发现算法	所有实体	
自动化事实核查算法	所有实体	

评估为高风险的情况包括：一类新闻媒体使用的或大规模媒体算法服务提供者的新闻推荐算法；一类新闻媒体使用的或大规模媒体算法服务提供者的自动化新闻写作算法；所有实体使用或提供的深度合成算法。

评估为中风险的情况包括：二类新闻媒体使用的或小规模媒体算法服务提供者的新闻推荐算法；二类新闻媒体使用的或小规模媒体算法服务提供者的自动化新闻写作算法。

评估为低风险的情况包括：三类新闻媒体使用的新闻推荐算法；所有实体使用或提供的新闻线索发现算法；所有实体使用或提供的自动化事实核查算法。

受算法技术、风险应对能力等诸多因素影响，新闻生产算法风险的级别需动态调整。第一，特定算法技术发展初期，由于性能和规制措施不足，可定为高风险。经一段时间发展，待性能、规制措施有效时，可以调至中、低风险。例如，深度合成目前应定级为高风险，后续可依据实际情况降级。第二，媒体算法服务提供者和算法服务使用者的用户规模变动也可能会改变影响等级，相关评估应视情况以年为单位做周期性的动态调整。

3.基于风险等级确定责任义务

不同的风险等级确定不同的责任义务。在确定落实的主体上,笔者认为可操作的思路是将算法服务提供者作为责任义务主体。一般认为,算法开发者是算法风险的技术源头,在现有研究中,也有观点认为算法开发者和算法使用者有算法影响评估的责任义务,然而这种思路的可行性值得商榷。

一方面,算法开发者可能是临时组成的团队,作为主体并不稳定;另一方面,算法开发者也是按照委托方的意图设计算法,一旦出现问题,算法开发者可能把责任推给委托方。算法服务使用者在引进、使用算法服务时,可能没有专业能力对算法影响进行评估。因此,承担责任义务的主体应当是连接算法开发者和算法服务使用者的算法服务提供者。如果无法履行算法影响评估的责任义务,算法服务提供者就不得提供算法服务。

算法责任义务包括算法影响评估报告、算法备案、算法审计三个方面。

(1)算法影响评估报告。算法影响评估报告可分为技术评估、风险等级评估、风险缓解(risk mitigation)措施评估。其中,技术评估主要评估算法的可靠性,可参考中国电子技术标准化研究院发布的《人工智能深度学习算法评估规范》;风险等级评估需要基于前文所述因素确定算法风险等级;风险缓解措施评估则是针对可能存在的算法失效、具体的算法风险问题、影响对象的权益侵犯进行客观描述,并就如何努力缓解或避免风险进行说明。例如,传统的新闻实践强调依赖可靠来源和确保信息准确性的重要性。然而,生成式人工智能在报道过程中内容的来源和可靠性方面存在不确定性[15]。面对这种情况,评估时应当陈述风险是如何缓解的,是清晰地标注了信源便于记者核查并设置了提醒功能,还是系统所依赖的信源都是来自可靠网站的?如果是这样,可靠来源需要一一列出。

(2)算法备案。算法备案是向监管部门备案算法部署情况,备案内容包括:算法基本情况信息,主要包括算法类型、算法名称、算法数据、算法模型、算法策略和算法风险与防范机制等信息;算法安全自评估情况;算法安全合规内部制度建设[26];算法影响评估报告。对涉及商业秘密保护的信息,只留存算法监管机构内部系统即可,反之需要信息公开,接受社会监督。由于新闻生产算法基本上属于具有舆论属性或社会动员能力算法,所以基本都要进行算法备案。

(3)算法审计。算法审计(algorithmic auditing)是对算法的安全性、合法性和伦理进行评估、缓解和确保的研究和实践,涉及人工智能的公平性、可解释性、鲁棒性、隐私等方面的研究[27],是一种证明算法存在偏差的研究方法。算法审计正成为倒逼各类与公共利益相关的平台进行算法改良、提升社会责任感的重要手段。高风险等级需要对算法进行强制审计。欧洲法律研究会(ELI)制定的《欧洲公共管理算法决策影响评估示范规则》(*Model Rules on Impact Assessment of Algorithmic Decision-Making Systems Used by Public Administration*)明确指出,算法影响评估为高风险的算法必须完成算法审计,否则使用算法决策系统是违法的[28]。

算法审计依照用途可以分为三类:技术审计(technical audit)、治理审计(governance audit)、实证审计(empirical audit)。

技术审计允许"深入了解"以找出系统中可能存在问题的地方。技术审计可以探索系统的内部机制,看数据、源代码或方法是否存在问题,包括7个方面的内容:评价系统输入,如测试数据是否平衡且高质量;评估模型的开发,研究用于训练算法的优化标准(例如损失函数)以及用于在开发过程中评估算法的性能指标;模型的构建、训练和测试方式;与开发人员访谈,以了解系统的工作原理;审计用于控制风险的措施,例如减轻偏见的技术控制;通过测试数据集并进行模拟对模型进行压力测试;检查代码等[29]。技术审计由于涉及算法系统过多细节问题,通常由算法开发者自我审计。

治理审计是广泛评估部署算法系统的组织是否具有管理其使用的适当政策,以及在围绕系统设计和实施的程序和文件中是否遵循了良好的做法,包括算法的透明性、可解释性,人类监督的水平、效率和有效性,主要涉及合规评估[29]。

实证审计旨在通过评估系统的输入和输出来衡量使用算法系统的效果。在可以访问系统的情况下,如通过API或第三方沙箱(third-party sandbox),审计者使用测试数据集时测试系统生成的输出,但不审计系统本身的工作。实证审计可以评估算法系统的输出是否存在问题,但通常不会揭示这些问题存在的原因或如何解决这些问题[29]。

对于低风险算法,算法服务提供者需要开展治理审计;对于中风险算法,算法服务提供者需要开展治理审计和技术审计;对于高风险算法,算法服务提供者需要开展治理审计、技术审计、实证审计。

针对低风险算法,治理审计可以由算法开发者自行审计,或由算法服务提供者委托第三方审计;针对中风险算法,治理审计和技术审计可以由算法开发者自行审计,或由算法服务提供者委托第三方审计;针对高风险算法,技术审计可以由算法开发者自行审计,治理审计和实证审计必须委托第三方审计。

此外,针对高风险算法,算法服务提供者需要提供三类数据访问权限用于算法审计需要,即一般数据访问权限、定制数据访问权限和监管数据访问权限[30]。一般数据访问权限是算法平台向社会提供关于算法系统运行的最低限度数据访问权限,这种访问权限提供的是通用的、无差别的简单数据集和相关介绍材料,开放这种权限便于社会随时进行监督。定制数据访问权限是针对某个具体风险审计任务而提供的数据权限,这种数据可以是最少化的、匿名的、去标识化和加密的。监管数据访问权限则是出于监管需要而提供的数据粒度上更为细致的复杂数据集,需要算法服务提供者提供相应的答辩材料。

由于算法系统是动态和变化的,有必要定时、定向开展算法审计工作。“定时”是以“年”作为审计周期,“定向”是在算法技术发生重大变化或某一问题集中显现时,针对特定问题开展算法审计。定时、定向审计结果既需要内部存档留作算法优化的设计依据,也需向相关部门报备。对已经部署有高级别风险的算法系统,应定期审核、评估、验证算法机制机理、模型、数据和应用结果,接受社会监管。

结　语

随着算法日益深入地嵌入新闻生产各个环节,算法风险也将更加突出。新闻业的公共属性决定了新闻生产算法风险治理的精准性和迫切性。《关于加快推进媒体深度融合发展的意见》指出,要以先进技术引领驱动融合发展,并提到要“用好”人工智能。善用人工智能的前提是新闻业对新闻生产算法风险有清醒的认识并能进行有效评估。算法影响评估机制可以有效协调创新与风险的冲突、降低治理成本、实现精准治理,让新闻生产更加从容地接纳和利用算法。在具体机制的构建上,确定风险等级是核心问题,风险等级是情境性的、相对的,这意味算法影响评估机制应是复杂的、动态的,如此才能务实可行。本文对新闻生产算法影响评估机制进行了初步探索,提供了构建的思路,提出的方案还有进一步细化、论证的空间,未来还应当深入研究算法技术类型,并将算法透明、算法可解释、算法问责、算法公平等伦理原则纳入其中,构建更立体、全面、细致的算法影响评估机制。

注释

①调研受访者 1 为媒体算法服务供应商(2022 年 12 月 4 日访谈)。

②调研受访者 2 为平台媒体个性化广告推荐程序员(2022 年 12 月 5 日访谈)。

参考文献

[1]COOLS H, DIAKOPOULOS N. Towards guidelines for guidelines on the use of generative AI in newsrooms[EB/OL].(2023-07-10)[2023-08-19].https://generative-ai-newsroom.com/towards-guidelines-for-guidelines-on-the-use-of-generative-ai-in-newsrooms-55b0c2c1d960.

[2]BAUDERD. AP, other news organizations develop standards for use of artificial intelligence in newsrooms[EB/OL].(2023-08-17)[2023-08-19].https://apnews.com/article/artificial-intelligence-guidelines-ap-news-532b417395df6a9e2aed57fd63ad416a.

[3]互联网信息服务算法推荐管理规定[EB/OL].(2022-01-04)[2023-08-19].http://www.cac.gov.cn/2022-01/04/c_1642894606364259.htm.

[4]互联网信息服务深度合成管理规定[EB/OL].(2022-12-11)[2023-08-19].http://www.cac.gov.cn/2022-12/11/c_1672221949354811.htm.

[5]张恩典.算法影响评估制度的反思与建构[J].电子政务,2021(11):57-68.

[6]KUNER C, CATE F H, MILLARD C, et al. Risk management in data protection[J].International Data Privacy Law, 2015(2): 95-98.

[7]范为.大数据时代个人信息保护的路径重构[J].环球法律评论,2016(5):92-115.

[8]朱成光.以风控为导向的系统导入与制度整合[EB/OL].(2022-12-27)[2023-08-19].https://www.chinatimes.com/newspapers/20221227000126-260210?chdtv.

[9]美国算法问责法案(2022)[EB/OL].吴建昊,译.(2022-02-20)[2023-08-19].https://mp.weixin.qq.com/s/y_9IQMw4jOmeGd-fGneAzw.

[10]谈毅.新兴科技领域风险治理应当加强公众参与[J].国家治理,2020(35):32-36.

[11]宁宣凤,吴涵,吴舸,等.路未央,花已遍芳:欧盟《人工智能法案》主要监管及激励措施评述[EB/OL].(2023-08-18)[2023-08-19].https://mp.weixin.qq.com/s/VMSAOuWYlg4U8_LcJocIGw.

[12]PORLEZZA C. Promoting responsible AI: a European perspective on the governance of artificial intelligence in media and journalism[J].Communications,2023(3): 370-394.

[13]European Broadcasting Union. AI act: high-risk AI systems need more nuance[EB/OL].(2022-09-09)[2023-08-19].https://www.ebu.ch/news/2022/09/ai-act-high-risk-ai-systems-need-more-nuance.

[14]JONES E. Inform, educate, entertain ... and recommend? [EB/

OL]. (2022-11-24)[2023-08-19].https://www.adalovelaceinstitute.org/report/inform-educate-entertain-recommend/.

[15]COOLS H, DIAKOPOULOS N. Writing guidelines for the role of AI in your newsroom? Here are some, er, guidelines for that[EB/OL]. (2023-07-11)[2023-08-19]. https://www.niemanlab.org/2023/07/writing-guidelines-for-the-role-of-ai-in-your-newsroom-here-are-some-er-guidelines-for-that/.

[16]张晒.风险分配何以公正:基于新冠肺炎疫情的哲学审思[J].北京理工大学学报(社会科学版),2020(3):57-64.

[17]于骐鸣.后现代网络技术哲学思想研究[M].武汉:华中科技大学出版社,2019:64.

[18]PREIMESBERGER C J. Why 2022 is only the beginning for AI regulation[EB/OL]. (2022-03-21)[2023-08-19].https://venturebeat.com/ai/why-2022-is-only-the-beginning-for-ai-regulation/.

[19]FARHI P. A news site used AI to write articles. It was a journalistic disaster[EB/OL].(2023-01-17)[2023-08-19].https://www.washingtonpost.com/media/2023/01/17/cnet-ai-articles-journalism-corrections/.

[20]卫才胜.技术的政治:温纳技术政治哲学思想研究[M].武汉:武汉大学出版社,2017:142.

[21]DOGRUEL L, MASUR P, JOECKEL S. Development and validation of an algorithm literacy scale for internet users[J]. Communication Methods and Measures, 2022(2):115-133.

[22]JONES B, JONES R, LUGER E. AI 'Everywhere andnowhere': addressing the AI intelligibility problem in public service journalism [J]. Digital Journalism, 2022(10):1731-1755.

[23]关于加强互联网信息服务算法综合治理的指导意见[EB/OL]. (2021-09-29)[2023-08-19].http://www.cac.gov.cn/2021-09/29/c_1634507915623047.htm.

[24]第一! 百度智能云领跑 2020 年中国智能媒体方案市场[EB/OL]. (2021-11-19)[2023-08-19].https://mp.weixin.qq.com/s/xGjGesyA1fVtcL4GTBPI5Q.

[25]中国新闻事业发展报告(2022 年发布)[EB/OL]. (2022-05-16)[2023-08-19]. http://www.zgjx.cn/2022-05/16/c_1310592108.htm.

[26]张吉豫.论算法备案制度[J].东方法学,2023(2):86-98.

[27]KOSHIYAMA A, KAZIM E, TRELEAVEN P, et al. Towards algorithm auditing: a survey on managing legal, ethical and technological risks of AI, ML and associated algorithms[EB/OL].(2021-02-15)[2023-08-19]. https://papers.ssrn.com/sol3/papers.cfm? abstract_id=3778998.

[28]Europen Law Institute. Model rules on impact assessment of algorithmic decision-making systems used by public administration[EB/OL].(2022-08-17)[2023-08-19].https://www.europeanlawinstitute.eu/fileadmin/user_upload/p_eli/Publications/ELI_Model_Rules_on_Impact_Assessment_of_ADMSs_Used_by_Public_Administration.pdf.

[29]Digital Regulation Cooperation Forum. Auditing algorithms: the existing landscape, role of regulators and future outlook[EB/OL]. (2022-09-23)[2023-08-19]. 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.

[30]张超.资讯类推荐算法的算法审计路径、伦理与可审计机制[J].中国出版,2021(7):31-35.

Construction of Algorithm Impact Assessment Mechanism in News Production

Zhang Chao Chen Sha

Abstract:The algorithm in the era of intelligent media has been widely embedded in the whole process of news production, and the various risks brought by it deserve attention. At the current national level, algorithmic risk governance focuses on recommendation algorithms and deep synthesis algorithms with public opinion attributes or social mobilization capabilities, but does not include other algorithms related to news production, forming a blind area of algorithmic risk governance. News production is closely related to public interests, so it is imperative to build a special algorithmic impact assessment mechanism for news production. The algorithm impact assessment has the advantages of subdivision of governance object, advance of governance node and dynamic assessment method. For news production, algorithm impact assessment can improve the pertinence of news production algorithm risk governance, fairly allocate risk responsibilities, effectively protect users' rights and interests, and improve the algorithmic literacy of news practitioners. The impact assessment of news production algorithm should identify the risk category according to the risk impact object of the algorithm, determine the risk level by combining multiple factors, and confirm the responsibility and obligation based on the risk level to build a specific evaluation mechanism.

Key words:algorithm impact assessment; algorithmic risk; risk-based approach; algorithmic literacy

责任编辑:沐　紫