【新闻与传播】

自动化事实核查的算法逻辑、内生性风险及其规避*

张　超

摘　要:智媒时代网络信息内容生态治理需要智能化手段,自动化事实核查是打击谎言和假新闻的算法治理手段之一。自动化事实核查的算法逻辑各不相同,总体可分五类:基于“匹配”的逻辑、基于“信源信度”的逻辑、基于“关系”的逻辑、基于“缺陷”的逻辑和基于“区块链”的逻辑。因“算法形式主义”引发的内生性风险会损害自动化事实核查的可信度。规避内生性风险,在技术层面,应优化设计,减少源数据的偏差;在把关层面,应采用“算法+事实核查员”的人机协同模式;在利益相关者层面,应组建事实核查网络;在伦理层面,应强化透明和更正原则。

关键词:自动化事实核查;算法形式主义;内生性风险;透明性

中图分类号:G206　　**文献标识码**:A　　**文章编号**:1003-0751(2022)02-0166-07

当前,全球网络信息内容生态治理的棘手问题是假新闻横行、不实言论混淆视听。智媒时代虚假内容的生产与传播也具有“智”的特征:造假手段的智能化(如深度伪造)和传播扩散的“拟人化”(如机器人账号)。2020 年,全球 84 个国家有 304 个事实核查项目①,项目数比 2019 年增加了近 100 个,但与全球每天产生的海量虚假内容相比远远不足。造假技术的智能化迫切需要事实核查的智能化,自动化事实核查(Automated Fact-Checking,简称 AFC)应运而生。

自动化事实核查用数字工具来识别、验证和回应误导性“陈述”(claims)。②它的最大优点在于处理速度快,如果用人工阅读文稿和监测电视寻找事实性“陈述”需要数小时,用自动化事实核查工具 ClaimBuster 只需要几秒钟。③自动化事实核查还可以规避某些公众的“批评”:传统的人工事实核查被视为有“党派偏见”。在美国,就有批评者认为,事实核查网站 PolitiFact 在选择核查事实上存在党派偏见,更容易选择共和党的失实言论作为核查对象,尽管民主党人与共和党人的核查文章总数相当,但出自共和党人的言论更容易被标注为“失实”或“荒谬”。④以“技术中立”面貌出现的自动化事实核查可以赋予事实核查客观、中立的形象。自动事实核查在核查“陈述”方面有一定的成效。联合国的研究报告显示,一些新闻媒体和事实核查机构利用自动化工具,加快了新冠肺炎病毒相关的事实核查速度。⑤

国内学界对事实核查的关注多聚焦于传统人工事实核查的流程、伦理、效果和个案研究,对自动化事实核查的技术逻辑及其蕴含的风险缺少关照。自动化事实核查的核心是算法驱动,是打击谎言和假新闻的算法治理(algorithmic governance)手段之一,但核查过程并非人工核查在技术上的“翻版”。从技术逻辑上看,人工事实核查与自动化事实核查并非“等同”关系,人工核查所采用的“匹配”逻辑在自动化事实核查中只是技术方法之一。计算科学家秉持的“算法形式主义”(algorithmic formalism)虽然让算法解决问题的思路变得清晰,让复杂的问题易于处理⑥,却让自动化事实核查蕴含技术风险。

基于此,本文拟从算法技术的内部视角深入分

收稿日期:2021-06-17

*基金项目:国家社会科学基金一般项目“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副院长、教授、博士生导师,广播电视学博士(威海　264209)。

析自动化事实核查的算法逻辑、内生性风险，并据此提出规避风险的举措。

一、自动化事实核查的算法逻辑

算法是为解决一个特定的问题所采取的确定的有限步骤。⑦解决一个问题可以有多个算法设计，算法性能可能存有差异。在自动化事实核查问题的解决上，算法逻辑也各不相同。

1.基于“匹配”的逻辑

自动化事实核查中基于“匹配”的逻辑源于人工事实核查。基于“匹配”的算法逻辑是以特定的语料库为基础的，当“陈述”出现时，系统自动将其与语料库的内容进行对比，如果匹配成功，则完成事实核查。事实核查的对象是“陈述”，而不是观点。应用这种逻辑的前提是要有一个“比对库”——基于以往已被验证事实的数据库。这是一种典型的将现实思路模拟为技术思路的核查方法。

2020 年新冠肺炎疫情期间，“国际事实核查网络”（International Fact-Checking Network）建立了一个事实核查数据库，包含超过 40 种语言的 7000 多个已核查“陈述”，聊天机器人 WhatsApp Chatbot 能够从这个数据库中找到匹配的“陈述”对用户提出的关键字请求进行核查回应。⑧

人工智能程序 Squash 可以实时匹配在事实核查系统 ClaimReview 中已有的事实核查与现场演讲者的“陈述”。⑨Squash 可以将政客的言论转录成可搜索的文本以便查找匹配项，几秒钟内将相关的事实核查显示在观众的屏幕上。⑩“匹配”逻辑之所以可行，是因为在各种场合、平台中的失实或误导性“陈述”都是重复的说法，而这些“陈述”可能已被事实核查过了。类似的算法系统还有《华盛顿邮报》的“吐真者”（Truthteller）和 FullFact 的“实时平台”（Live platform）。

基于“匹配”逻辑的事实核查方法看似简单，想做好也非易事。这种方法需要前期积累大量的已验证事实数据，由于事实核查的语料库毕竟是有限的、时滞的，新的语料会不断涌现，因而需要持续更新。例如，事实核查初创公司 Logically 计划在一年内积累 10 亿个事实，以便能够自动核查尽可能多的“陈述”。⑪

2.基于“信源信度”的逻辑

如果不分析“陈述”内容本身的真假，如何预测“陈述”的真假？基于“信源信度”的逻辑是一种典型的计算思维，它通过对信源可信度的判断推论出“陈述”的可信与否。

针对信源的信度，有三种假设。第一种假设是机器人账户容易传播假消息，只要识别出信源是机器人账户，就可以判定为假。针对 Twitter 的模拟实验发现，如果将机器人内容排除在外，低可信度文章的转发总量会减少 70%。⑫目前，识别机器人账户主要采用机器学习的方法。印第安纳大学开发的 Botometer 经过训练可根据 3 万个账户的数据集中模式识别机器人行为。它为每个账户读取超过 1000 个不同的特征，然后给账户分配一个 0 到 1 之间的分数，分数越高，账户是机器人的可能性越大。⑬一旦账户的分数突破设定，则会被判定为机器人账户，其散布的言论将被视作谣言。

第二种是基于信源过往的内容发布情况。如果信源之前发布过假新闻，则意味着它可能再次发布，会被判定为低可信度信源。麻省理工学院的研究团队通过判断整个新闻网站是否经常呈现准确或错误的信息来判断信源可信度。该团队通过机器学习算法运行“媒体偏见/事实核查”（Media Bias/Fact Check）网站中有关新闻来源的评估数据，根据网站内容的文本、句法和语义分析对 1000 多个新闻网站进行分类，重点关注结构、情感、参与度、主题、复杂性、偏见和道德等特征，只需分析一个网站的 150 篇文章就能评估该网站是否可信。⑭

第三种方法是假设自动文本合成器（automatic text generators）生成的内容都是不可信的。在全球互联网的虚假内容产制中，自动文本合成器扮演了重要角色。OpenAI（开放人工智能）组织开发的 GPT-2 语言模型被用于写作辅助、摘要等领域。有研究显示，超过 70% 的人认为这些文本就像《纽约时报》的文章一样可信。⑮由于该技术也被滥用于快速产生大量虚假信息，研究人员开发了识别这类自动文本的方法。⑯

基于“信源信度”的逻辑用整体封锁信源的方式杜绝假新闻，看似简单粗暴，但对付自动化的假新闻批量生产确实有效。

3.基于“关系”的逻辑

将“陈述”视为知识，如果为真，则一定有支持它的相关知识，反之则很少或没有。自动化事实核查的另一个逻辑在于“关系”，即寻找某一“陈述”在

知识图谱中与其他知识的关系。

知识图谱是结构化的语义知识库,用于以符号形式描述物理世界中的概念及其相互关系。其基本组成单位是“实体—关系—实体”三元组,以及实体及其相关属性—值对(Attribute-Value Pair),实体间通过关系相互联结,构成网状的知识结构。[17]有研究通过挖掘知识图上的连接模式来计算“陈述”的支持度,用维基百科的大规模知识图来计算简单“陈述”的支持度。通过在知识图上适当定义的语义邻近度量下找到概念节点之间的最短路径,这种方法近似人类事实检查时的复杂判断。[18]有研究依靠直接事实逻辑的核查方法——“立场检测”(stance detection),使用深度学习(deep learning)算法来确定帖子或故事中的声明是否得到同一主题的其他帖子和故事的支持。滑铁卢大学开发的系统准确率为90%。随着时间的推移,当显示新的“陈述”时,系统还学会了自己确定支持还是不支持。[19]

4.基于“缺陷”的逻辑

人工智能技术的发展让“深度合成”游走于“合成”和“伪造”的两端。深度伪造成为事实核查新的核查对象。面对这种新的、足以以假乱真的文本形态,传统的人工识别难以应对。如果从技术的角度去解决,则有很多突破口。因为无论是合成还是伪造,从数据上看都会有“缺陷”,如果能找出这类视频在“缺陷”上的数据特征,则可以有效识别真假。

深度伪造的缺陷之一在于体态语的不自然。现有技术通过研究说话者嘴唇动作、语言模式和手势之间的关联性来判断是否属于深度伪造。[20]以人物合成视频为例,这种合成图像算法的缺陷是“眨眼”频率比真人少得多,辨别真假视频的原理就是用机器学习来检查视频中的“眨眼”频率,准确率超过95%。[21]

深度伪造的缺陷之二在于凡是伪造都会有“痕迹”。微软视频认证器(Video Authenticator)通过在视频播放的每个帧与原图进行实时对比,给出数据分析,可以检测出人眼可能无法察觉到的合成修改痕迹。[22]由 Jigsaw 公司新闻工作者开发的工具 Assembler,可以通过五个检测器进行内容分析来验证内容的真实性,包括检测图案和颜色、复制和粘贴区域的异常,以及 Deepfake 算法的已知特征。[23]

5.基于“区块链”的逻辑

基于“区块链”的逻辑是依靠文本上的“元数据”(如时间、地点、作者以及所有编辑和发布的信息)来判断文本的来源及其是否被篡改。基于“区块链”的逻辑需要从信息文本(图片、文章、图像等)的源头开始布局,如此才能真正发挥作用。当前,全球部分主流媒体已经开始布局区块链,通过其核查不实信息。

意大利安莎通讯社与安永咨询公司合作开发了一款基于区块链技术的新闻跟踪系统。用户可以通过 ANSAcheck 的新闻跟踪标签应用,查看出现在安沙通讯社平台或是分发给其他刊物或第三方平台的新闻的来源。[24]《纽约时报》正在研发区块链项目“新闻出处追溯”(The News Provenance Project),用来创建和共享新闻图片的“元数据”。媒体和用户可以判断出该图片是否经过 PS 等人为修饰,进而判断相关资讯是否是假新闻。

2019 年,Adobe 公司与《纽约时报》、Twitter 合作,通过“内容真实性计划”(Content Authenticity Initiative,简称 CAI)打击虚假数字新闻内容,该系统通过标记新闻内容的完整历史和记录(包括位置、时间戳以及是否被编辑过等)验证真假,这有助打击网上虚假讯息和假照片的传播。[25]

二、自动化事实核查的内生性风险

技术风险按生成方式可以分为外生性风险和内生性风险。外生性风险是由技术之外的因素引发的风险。例如,使用者对于技术的误用、误解和滥用。内生性风险是与技术设计本身直接相关的各类因素引发的风险。

自动化事实核查的内生性风险主要是由“算法形式主义”引起的,形式主义是算法的核心要义,意味着坚持规定的形式和规则,算法需要输入、输出和目标的明确数学表达[26],并将解决方案公式化、抽象化和规则化,整个过程是一个转译的过程,而转译意味着偏差。有专家认为,计算机科学家应该关注模型的抽象和世界的复杂性之间的鸿沟。[27]在弱人工智能阶段,算法的处理还不具备处理复杂问题的能力,将复杂问题简单化的转译过程决定了偏差出现的必然性,加之事实核查是一种动态的、语境性的、复杂的工作,其内生性风险不可避免。自动化事实核查的内生性风险主要有误解风险和误判风险。

1.误解风险

误解风险是指由于事实核查系统无法准确“理

解”人类语言而引发对“陈述”的误解。误解风险的产生是人工智能发展的阶段性问题。目前还处于弱人工智能阶段,算法能够从事的是简单的、重复的事实核查,变通性较差。

在基于“匹配”的算法逻辑中,匹配行为是一种机械地匹配,没有弹性空间。而语言表达是有弹性的,尤其是模糊语言(如“很多人”)在日常生活中很常用,但是对于算法而言,这种“模糊”会让其“迷茫”。因此,完全自动化的事实核查局限在统计信息这一狭窄领域。[28]

简单的“复制和粘贴”重复,相对容易核查,但更多情况下,被核查的“陈述”是原来释义的变体。[29]对人来说很简单的“相似表达”问题,对算法而言是一个难题。如果机械地用单词一一匹配就会出错。[30]例如,特朗普可能会提出一个有关太空旅行的“陈述”,但自动事实核查平台 Squash 会将其与之前有关修建道路的官僚主义作风的事实核查进行错误匹配。[31]

被寄予期望的自然语言处理(Natural Language Processing,简称 NLP)技术虽然能可靠地捕捉到一个表述的相近变体,但释义会是一个相当大的挑战,存在“查全率”(recall)和“查准率”(precision)之间的矛盾。[32]

在锚定语言含义的因素中,语境是至关重要的,但也是算法无能为力的。杜克记者实验室的首席技术专家克里斯托弗·格斯表示,当前的技术通常无法使计算机理解政客说话的方式和背景。[33]当然,即便是完全匹配成功,也并不意味着核查结果就一定正确。例如,有些政客用数据说话,数据是真实的,但是数据所处的时间段是精心选择的,结论以偏概全并不可靠,然而算法可能会判其为真。

2.误判风险

基于“信源信度”的逻辑、基于“关系”的逻辑、基于“缺陷”的逻辑,由于解决问题的思路不直接涉及被核查的内容本身,会产生误判风险。

在基于“信源信度”的逻辑中,识别内容是否是由机器生成的,这种方法存在漏洞,即使生成的文本包含了真实的事实,但这种方法会错误地假设所有机器生成的文本都是自动错误的。[34]例如,News-Cracker 在核查 BuzzFeed 网站的一篇报道时,将一篇原本客观的报道认定为有偏见,因为文中引用的用户推文被检测为“许多陈述无法得到验证”,从而对整个网站的可信度产生了影响。[35]

在基于“关系”的逻辑中,在核实新的和未经检查的主张时,自动化还不能很好地发挥作用,部分原因是自动化事实核查需要权威的数据,这些数据往往无法获得。[36]如“立场检测”,假如在数据库中没有包含支持某一“陈述”的帖子或报道,一个真实的“陈述”则很可能被判定为假。

在基于“缺陷”的逻辑中,通过寻找视频在生物特征上的缺陷的模式,只能在一定时期、一定阶段有效,因为深度伪造技术正在完善,生物特征测试越来越难以发挥作用。

在基于“区块链”的逻辑中,能够进行核查的前提是图片本身具有可被识别“身份”的元数据,如果缺少这种数据,则无法识别。因为发布在社交媒体上的图片,在上传时元数据会被剥离。[37]

三、自动化事实核查内生性风险的规避

在弱人工智能时代,自动化事实核查不可避免地存在缺陷。作为一种信息纠错机制,自动化事实核查内生性风险的最大后果在于损害事实核查与公众间的信任关系。如何规避内生性风险的产生,不仅是技术问题,还需要多措并举,在技术、把关、利益相关者和伦理等层面构筑立体的防范体系。

1.技术层面:优化设计,减少源数据的偏差

现有的误判和误报风险很大程度上是由算法偏差造成的。算法偏差的形成主要来源于算法设计和源数据。在算法设计上,复杂的问题需要复杂的算法进行解决,需要算法设计者在转化现实问题时考虑到算法设计的复杂性和情境性,不能迷失在“算法形式主义”的框架中。因为形式知识(formal knowledge)需要缩小视野,“在社会工程设计中编码的形式秩序不可避免地遗漏了对其实际功能至关重要的元素”[38]。只有那些在算法语言中可读的考虑因素,才被认为是重要的设计和评估考虑因素。[39]

在源数据方面,无论是基于哪种算法逻辑,自动化事实核查都需要大量源数据作为支撑。自动事实核查最重要的优先事项是确保事实核查者可以依赖的各种来源数据是计算机可以使用的结构化数据。[40]

对于基于“匹配”逻辑的自动化事实核查,核查者所建立的比对数据库至关重要。“如果数据来源分别是无党派的国会预算办公室和保守派的‘美国

人支持税制改革',系统得出的税收增长分析可能截然不同。"[41]英国事实核查机构 Full Fact 与英国国家统计局合作,用结构化数据作为复杂事实核查的部分依据。[42]阿根廷事实核查机构 Chequeado 的自动核查工具 Chequeabot 目前虽有 1000 个事实核查材料可供匹配,但缺少政府官方数据,该机构打算将高校、智库等机构纳入合作渠道,以弥补其缺陷。[43]统计分析不仅需要有权威的数据来源,还要提供结构化的数据,对于基于其他算法逻辑,尤其运用了机器学习、深度学习技术的自动化事实核查,训练数据集的质量则更为重要,因为前期训练数据有偏差,最终会导致系统习得的标准有偏差。例如,在针对图像的自动化事实核查中,许多训练集中的数据由计算机生成,或是从极为有限的公开图像中提取,因而有研究者建议,应与路透社、法新社或美联社进行数据库合作,用现实的新闻图像进行训练,这会在很大程度上提升算法的可靠度。[44]

2.把关层面:"算法+事实核查员"的人机协同模式

发布不正确的事实核查可能严重损害事实核查组织的声誉。[45]自动化事实核查的内生性风险使"人"在核查中的主体性凸显。为了避免自动化事实核查产生批量的、系统性的、后果严重的错误,将事实核查员"嵌入"各个把关环节尤为重要,由此形成"人在回路中"(human-in-the-loop)。例如,在选题环节,算法识别出值得核查的线索后,在进入自动化核查前,应当由事实核查员介入,判断哪些信息真正值得核查。因为并不是所有的可疑信息都值得核查,这里的"值得"取决于公共利益和社会关注度。在核查环节,系统核查结束后,也应该采用人工的方式进行二次审核。为了降低 Squash 的风险,技术团队在此基础上构建了一个新界面 Gardener,Squash 需要为其提供三个潜在的匹配事实核查结果,让事实核查员来选择呈现哪一个结果。[46]

3.利益相关者层面:组建事实核查网络

从全球范围看,事实核查是一种公益性的新闻事业。在人工事实核查方面,全球范围有多个项目进行合作。新冠肺炎疫情期间,"国际事实检查网络"组织了"新冠肺炎病毒事实联盟",这个联盟汇集了 70 个国家的 100 多名事实检查员来更新关于新冠肺炎疾病的虚假信息的数据库。法国的 Cross-Check 与 34 个新闻机构及新闻专业的学生联手,对法国总统大选进行报道。2015 年成立的"初稿新闻"(First Draft News)是由媒体、大学、平台和公民组织组成的事实核查协作体,它向记者和公众免费培训相关技能。

目前,部分事实核查机构运用的人工智能技术都是开源的,例如谷歌用于自然语言处理的 BERT 模型被 FullFact 用于检测句子是否匹配。但在更多的领域,自动化事实核查的相关合作还有待进一步加强和深入。在源数据上,应与权威机构(尤其是主流媒体)合作获得更多高质量的训练数据。在算法设计上,使用者和研发者还缺少充分地协作和沟通。在定义任务、开发数据集以训练模型和开发自动化系统方面,研究人员和实践者之间缺乏沟通,而且算法模型应该是可解释的、不偏不倚的,更符合伦理考虑。[47]同时,对于已经开发的自动化事实核查算法系统,也倡导科研人员和公众用算法审计的方式发现设计中的偏差,更好地优化系统性能。

目前,大多非营利性事实核查机构附属于新闻媒体或非政府组织,资金来源是一个问题。虽然一些事实核查组织已经创造出相对低廉的自动化事实核查工具,但要开发、推进大规模的自动化事实核查系统需要社会各方的持续性支持,否则也走不长远。2016 年,核查目击媒体内容的 Reportedly 因失去母公司的资金资助而停止更新。[48]

4.伦理层面:强化透明和更正原则

当算法系统不能保证百分之百正确时,对公众诚实的态度非常重要。只有这样,公众才能给予自动化事实核查容错的空间。当然,对于系统开发者和使用者来说,也不能以此为借口经常出错。此外,自动化事实核查还要避免成为某些利益相关者的"工具",行事实核查之名,做偏见、误导之事。

所以,新闻伦理中的透明性原则在自动化事实核查中要格外凸显。例如,在"国际事实核查网络"的事实核查者应遵守的五个原则中,涉及透明性的就有三个原则:对来源透明度的承诺、对资金和组织透明度的承诺、对方法透明度的承诺。[49]对于自动化事实核查而言,应当告知公众核查者是算法系统而非人类,并提供算法系统的一般核查原理(包括训练数据的相关情况、步骤等)和准确率,提醒公众注意可能会出现不准确的情况。

当自动化事实核查发生错误时,需要及时启动更正机制,并在页面注明"更正"字样,向公众解释

内容的变化。[50] PolitiFact 规定，如果发现事实性错误，该文章页面会加上“更正”的标签，并以编者按的形式说明详情；如果是补充或更新，则加上“更新”的标签并予以说明；如果发生重大错误以至于要改变真实性等级，将再次召集编辑小组，重新撰写文章，并在文章开头对这种变动予以说明。[51] 在网站上设置专门栏目“更正与更新”（Corrections and Updates），对先前有问题的事实核查重新刊登，并解释错误之处以及是否影响对“陈述”的评级。Full Fact 网站也设有专门的更正日志（Corrections Log）页面，告知读者为什么更正以及更正的内容。

四、结语

自动化事实核查是网络信息内容生态治理的重要手段，不同的算法逻辑展示了从算法角度进行事实核查的不同路径。面对“算法形式主义”带来的技术局限，如何规避风险、提高自动化事实核查的效率和准确性，是用技术治理技术问题的“元命题”。当智媒时代算法被“赋魅”时，利益相关者应当提高算法素养为其“祛魅”，只有辩证、批判地看待算法在自动化事实核查中的逻辑和效用，我们才能让自动化事实核查为新闻业的事实核查事业“加分”，也才能更好地参与网络信息内容生态治理。

注释

①Mark Stencel, Joel Luther. *Fact-checking Count Tops 300 for the First Time*. https://reporterslab.org/fact-checking-count-tops-300-for-the-first-time/, 2020 年 10 月 13 日。②Julia Sittmann, Andrew Tompkins. *The Strengths and Weaknesses of Automated Fact-checking Tools*. https://www.dw.com/en/the-strengths-and-weaknesses-of-automated-fact-checking-tools/a-53956958, 2020 年 7 月 17 日。③⑩㉚Andrew Donohue. *Using Artificial Intelligence to Expand Fact-checking*. https://reporterslab.org/using-artificial-intelligence-to-expand-fact-checking/, 2019 年 9 月 16 日。④Eric Ostermeier. *Selection Bias? PolitiFact Rates Republican Statements as False at 3 Times the Rate of Democrats*. http://editions.lib.umn.edu/smartpolitics/2011/02/10/selection-bias-politifact-rate/, 2011 年 2 月 10 日。⑤[澳]朱莉·波塞蒂、[英]卡莉娜·邦切娃：《信息疫情：解密虚假新冠疫情信息》，李美译，联合国教科文组织网站，https://en.unesco.org/sites/default/files/disinfodemic_deciphering_covid19_disinformation_zh.pdf, 2020 年 11 月 26 日。⑥㉖㉗㊳㊴Ben Green, Salomé Viljoen. *Algorithmic Realism: Expanding the Boundaries of Algorithmic Thought*. Proceedings of the ACM Conference on Fairness, Accountability, and Transparency, 2020, pp.19-31.⑦王荣良：《计算思维教育》，上海科技教育出版社，2014 年，第 46 页。⑧Samuel Danzon-Chambaud. *Automated Fact-checking Can Catch Claims that Slip Past Human Checkers. Here are the Two Ways They Work*. https://www.poynter.org/fact-checking/2020/automated-fact-checking-can-catch-claims-that-slip-past-human-checkers-here-are-the-two-ways-they-work/, 2020 年 8 月 21 日。⑨㉝㊻Harrison Mantas. *How the Duke Reporters' Lab Used the Political Conventions to Perfect Its Automated Fact-checking Program*. https://www.poynter.org/fact-checking/2020/how-the-duke-reporters-lab-used-the-political-conventions-to-perfect-its-automated-fact-checking-program/, 2020 年 8 月 28 日。⑪Greg Noone. *AI vs Misinformation: Fighting Lies with Machines*. https://techmonitor.ai/technology/ai-and-automation/ai-vs-misinformation-fighting-lies-machines, 2021 年 6 月 14 日。⑫Tristan Greene. *Here's Why Low-credibility News Seems to Dominate Twitter*. https://thenextweb.com/artificial-intelligence/2018/11/27/heres-why-low-credibility-news-seems-to-dominate-twitter/, 2018 年 11 月 27 日。⑬John Gramlich. *Q&A: How Pew Research Center Identified Bots on Twitter*. https://www.pewresearch.org/fact-tank/2018/04/19/qa-how-pew-research-center-identified-bots-on-twitter/, 2018 年 4 月 19 日。⑭Kimberley Mok. *MIT Algorithm Sniffs Out Sites Dedicated to 'Fake News'*. https://thenewstack.io/mit-algorithm-sniffs-out-sites-dedicated-to-fake-news/, 2018 年 11 月 15 日。⑮郑意：《OpenAI 发布高级语言人工智能模型 GPT-2，参数达 7.74 亿个》，猎云网，https://www.lieyunwang.com/archives/458074, 2019 年 8 月 21 日。⑯Rachel Gordon. *Better Fact-checking for Fake News*. https://news.mit.edu/2019/better-fact-checking-fake-news-1017, 2019 年 10 月 17 日。⑰刘峤、李杨、段宏、刘瑶、秦志光：《知识图谱构建技术综述》，《计算机研究与发展》2016 年第 3 期。⑱Giovanni Luca Ciampaglia, Prashant Shiralkar, Luis M. Rocha, Johan Bollen, Filippo Menczer, Alessandro Flammini. Computational Fact Checking from Knowledge Networks. *PLoS One*, 2015, Vol. 10, No.6, DOI:10.1371/journal.pone.0128193.⑲Media Relations. *New Tool Uses AI to Flag Fake News for Media Fact-checkers*. https://uwaterloo.ca/news/news/new-tool-uses-ai-flag-fake-news-media-fact-checkers, 2019 年 12 月 16 日。⑳Gregg Rosenzweig. *How to Spot Deepfakes and Fake News: Tips and Detection Tools*. https://www.clearvoice.com/blog/how-to-spot-deepfakes/, 2020 年 2 月 10 日。㉑人工智能 T800：《Deepfake 机器学习算法能够伪造逼真视频》，搜狐网，https://www.sohu.com/a/272156100_100136762, 2018 年 10 月 30 日。㉒贝爽：《Deepfake 搅动政局？微软推出最新检测工具，助力美国大选免遭虚假信息威胁》，雷峰网，https://www.leiphone.com/category/industrynews/o7kUebDfvl1t4SMF.html, 2020 年 9 月 3 日。㉓Carole Meziat, Nour Amara, Pierre Caussanel. *Deep Dive into Deepfake — How to Face Increasingly Believable Fake News?*. https://www.riskinsight-wavestone.com/en/2020/05/deep-dive-into-deepfake-how-to-face-increasingly-believable-fake-news-2-2/, 2020 年 5 月 18 日。㉔《疫情时期 ANSA 利用区块链技术追溯新闻真实性》，搜狐网，https://www.sohu.com/a/386286947_100268681, 2020 年 4 月 8 日。㉕《Adobe 将为“P 图”推标签识别系统，打击虚假内容传播》，界面新闻，https://www.jiemian.com/article/4829714.html, 2020 年 8 月 15 日。㉘㉜叶珂：《自动化事实核查的前景及局限性》，《传媒评论》2018 年第 5 期。㉙㊺㊼Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, Giovan-

ni Da San Martino. *Automated Fact-Checking for Assisting Human Fact-Checkers. https://arxiv.org/abs/*2103.07769, 2021年5月22日。㉛ Bill Adair, Mark Stencel. *A Lesson in Automated Journalism: Bring Back the Humans*. https://www.niemanlab.org/2020/07/a-lesson-in-automated-journalism-bring-back-the-humans/, 2020年7月29日。㉞ Kimberley Mok. *MIT's New AI Tackles Loopholes in 'Fake News' Detection Tools*. https://thenewstack.io/mits-new-ai-tackles-loopholes-in-fake-news-detection-tools/, 2019年10月25日。㉟㊽杨丽萍:《欧美新闻事实核查技术应用及趋势》,《中国传媒科技》2018年第6期。㊱Julia Sittmann, Andrew Tompkins. *The Strengths and Weaknesses of Automated Fact-checking Tools*. https://www.dw.com/en/the-strengths-and-weaknesses-of-automated-fact-checking-tools/a-53956958, 2020年7月17日。㊲㊹T. J. Thomson, Daniel Angus, Paula Dootson, Edward Hurcombe, Adam Smith. *Visual Mis/disinformation in Journalism and Public Communications: Current Verification Practices, Challenges, and Future Opportunities, Journalism Practice*. 2020, DOI: 10.1080/17512786.2020.1832139.㊵Full Fact. *The State of Automated Factchecking*. https://fullfact.org/media/uploads/full_fact-the_state_of_automated_factchecking_aug_2016.pdf, 2016年8月17日。㊶David Holmes. *Washington Post's Truth Teller and the Future of Robots Doing Journalism*. https://pando.com/2013/01/29/washington-posts-truth-teller-and-the-future-of-robots-doing-journalism/, 2013年1月29日。㊷Matt Burgess. *Google is Helping Full Fact Create an Automated, Real-time Fact-checker*. https://www.wired.co.uk/article/automated-fact-checking-full-fact-google-funding, 2016年11月17日。㊸Daniel Funke. *In Argentina, Fact-checkers□Latest Hire is A Bot*. https://www.poynter.org/fact-checking/2018/in-argentina-fact-checkers-latest-hire-is-a-bot/, 2018年1月11日。㊾The International Fact-Checking Network. *International Fact-Checking Network Fact-checkers' Code of Principles*. https://www.poynter.org/ifcn-fact-checkers-code-of-principles/, 2019年4月26日。㊿FactCheck.org. *Our Process*. http://www.factcheck.org/our-process/, 2020年8月12日。(51)万小广:《"事实核查"类新闻初创项目的启示》,《传媒评论》2014年第11期。

责任编辑:沐　紫

Algorithmic Logic of Automated Fact-checking and the Avoidance of Its Endogenous Risks

Zhang Chao

Abstract: In the age of intelligent media, the governance of Internet information content also needs intelligent means. Automated fact-checking is one of the algorithmic governance means to combat lies and fake news. Algorithmic logics are also different in solving automated fact-checking problems. Generally, they can be divided into five kinds: logic based on "matching", logic based on "source reliability", logic based on "relationship", logic based on "defect" and logic based on "blockchain". The endogenous risks of "algorithmic formalism" can undermine the credibility of automated fact-checking. To avoid risks, we should optimize the design at the technical level and reduce the deviation of the source data. At the checking level, the man-machine collaboration mode of "algorithm+fact checker" should be adopted. We also need to establish a fact-checking network at the level of stakeholders and strengthen the principle of transparency and correction at the ethical level.

Key words: automated fact-checking; algorithmic formalism; endogenous risk; transparency