价值冲突与治理出路：虚假信息治理中的人工智能技术研究

管佻路　顾理平

（南京师范大学 新闻与传播学院，江苏 南京 210023）

【摘 要】人工智能技术不仅能将虚假信息的生产自动化，也能用于检测和甄别虚假信息。本文旨在探讨“用技术手段来解决技术隐患的快速方案是否有效以及如何使之有效”的问题。基于真实信息以及搜寻是有成本的这一前提，本文从信息搜寻的视角出发，解释人工智能在生产和核查虚假信息方面对用户信息搜寻成本以及搜寻环境的影响。尽管人工智能技术能够用于删除虚假账号和高效识别虚假信息，但这种技术性方案也存在包括提高用户搜寻负担，造成信息环境鸿沟等诸多长期风险。本文进而探讨了在“自动化”虚假信息泛滥的网络困境中，跨部门应如何在人工智能技术及算法的使用上进行规制与合作，以保障人们对于真实信息的搜寻。文章认为，在信息核查的实践中不能忽视人工智能的公共性特征，一个连结各相关方的第三方机构应被建构，专业事实核查机构和新闻记者也应被给予更多支持。

【关键词】人工智能；虚假信息；生产甄别；搜寻成本

【中图分类号】G210　　【文献标示码】A

一、问题的提出

虚假信息从来不是一个新问题，自人类社会诞生，口口相传的谣言就随之出现。但同时，虚假信息又与技术条件和时代背景紧密相关，因此其传播特征处于不断的变动中。近年来，虚假信息成为一个热门话题被旧事重提，一方面缘于互联网和社交媒体的流行，使它能被人们更频繁和更轻易地接触到，另一方面也因为虚假信息在一系列人类的重大事件中展现出的巨大杀伤力。例如，在新冠肺炎疫情中，大量有关疫情的虚假信息传播对于个体的精神状态造成伤害，影响了公众和个人的科学应对决策，甚至加快了疫情的传播进程（Tedros Adhanom Ghebreyesus，2020），因此被称为与传染病疫情相伴的“信息疫情”（infodemic）。在社交网站和传播技术飞速发展的今天，虚假信息

[作者简介] 管佻路，南京师范大学新闻与传播学院博士研究生。
顾理平，南京师范大学新闻与传播学院教授，中国新闻史学会媒介法规与伦理研究委员会会长。

的传播面向也与此前十分不同——虚假信息大多由真实的人类用户创作和传播的情形已然迅速改变，人工智能技术越来越直接和深入地参与虚假信息的生产和传播。人工智能算法能够快速生成大规模足以乱真的虚假信息和评价。有关美国最大的点评软件Yelp的研究显示，机器生成的虚假评论不仅难以被人类用户分辨，而且网站自身的过滤系统和剽窃检测软件都无法识别出真假（Yao et al.，2017）。

在社交网站上，人工智能技术能够在短时间内创建出数以百万的机器人账号以传播虚假信息。一项针对推特的实验研究显示，如果将机器人账号内容排除在外，虚假信息的转发总量会减少70%（Tristan，2018）。这些社交机器人由自动程序创建，通过高频率地发布和重复虚假信息，并使用相似的虚假内容回复真实用户的内容，达到操控社交对话的目的。由于普通用户倾向于相信自己看到次数多或分享人数多的内容，而鲜少质疑（如仔细查看账户的用户资料和过往发文内容），虚假的社交机器人账号得以凭借巨大的虚假信息数量占据互联网空间，并抢夺用户注意力。不仅如此，人工智能和算法还能够让虚假信息的生产更加机动和灵活。根据乔治城大学的一项研究报告，GPT-3（一种强大的人工智能系统）可实现重复叙述、阐述、操纵、说服，根据情境自主开发新叙述，以及精准定位新成员目标等常见的虚假信息宣传任务（Buchanan et al.，2021）。研究人员还发现，GPT-3在每项任务上都表现出色，几乎不需要人工参与就可以在短短几分钟内设计和制作出有针对性的虚假信息，模仿特定的写作风格，从而有效影响受众的态度。

与此相反，人工智能技术正在被越来越多地运用于虚假信息治理，主要集中于核查虚假信息内容生产者（如社交媒体上的虚假账号）和信息内容本身，即利用机器学习技术对虚假账号和信息进行识别，建立由人工智能驱动的防御和过滤系统来阻止其传播。根据核查系统的输入源和计算原理，人工智能新闻核查技术可以分为依托自然语言处理技术的新闻内容模型和关注社交特征和信号的社会情境模型（师文，2018）。例如，哈佛大学和MIT-IBM沃森实验室的研究者开发出的开放性工具GLTR就是基于内容的核查算法，可较为准确地识别出哪些文本更可能是由人工智能生成的。而对社交媒体平台上的虚假账号核查，则更多地使用基于社会情境的算法模型。如Facebook正在使用的新的机器学习系统“深层实体分类”（deep entity classification）不再孤立地查看单个账户资料，而是查看账户和页面间的联系，使用系统聚合指标来建立更详细和全面的账户统计信息。经过机器学习和神经网络训练后的用户数据模型能够更准确地将账户分类以鉴别账户真假。也正是得益于这些人工智能工具，Facebook在2020年删除了58亿个用于散播虚假信息和广告的虚假账户。[1]

如同一场军备竞赛，信息侦察技术的不断成熟又反过来令制假的水平更高超且更隐

蔽。一方面，虚假信息的表达途径不再拘泥于通过文本，而是扩充到了图像、声音和视频中。机器学习技术被滥用于生产虚假视频——深度伪造工具（Deepfake）掀起了一股“前所未有的虚假或诽谤性内容的浪潮”（Warner，2018），该技术能够在原本的图像和视频中的人脸上叠加进目标人物，合成极其逼真的全新视频。另一方面，利用人工智能制造虚假信息的门槛显著降低。哪怕没有专业的视频编辑技能，普通人也可以使用智能手机上的应用程序来制作简易的深度造假视频。技术制假呈现出产业化和规模化。在巨大利益的刺激下，利用人工智能生产虚假信息的网络变得愈发庞大且分工明确。

可以说，如今人工智能技术已经融入虚假信息的生成、传播再到治理的整个循环过程。目前已有研究主要集中在人工智能治理虚假信息的环节。研究包括从技术层面对算法治理进行分析，例如基于微博用户的用户特征分析，使用卷—长短期记忆网络模型实现对于微博谣言的早期检测（尹鹏博等，2020）；再如以复杂网络关系发展理论为基础，将谣言与真实信息进行信息轨迹聚类与隔离，提出了覆盖度和准确度，且性价比也更高的算法（王征，2019）。另有研究则专注于算法治理在特定事件中或特定类型的虚假信息中的作用机制，例如采用机器学习分类算法对健康类的谣言分享意愿进行建模与分析，为阻断健康谣言的传播提供算法干预建议（位志广等，2020）。还有研究将算法治理虚假信息（算法纠错）作为事实核查的类别之一，将其作为整体辟谣机制的一部分进行讨论。例如，基于微信朋友圈的研究发现，用户信息托付机制的复杂性，导致基于同一平台虚假信息的算法纠错和社会纠错的效果相差不大（杨洸等，2020）；再如基于香港本地虚假信息核查平台的实践经验，曾姿颖等（2021）提出了“人工核查—广众参与—人工智能”的协同核查模式。

本文以人工智能为核心，将其置于虚假信息传播的全过程中加以讨论。在一个流动的信息搜寻市场中，当真实信息成为用户需要付出搜寻成本才能得到的市场资源时，人工智能参与生产和治理虚假信息都成为必然发生的市场反应。然而，这种基于市场价值而出现的技术工具，能否“拿来主义”地对原有市场实现治理？现有实践证明，这种技术性方案并未达到预期目标。跟随这条逻辑思路，本文首先分析了人工智能在信息搜寻市场中对虚假信息生产和甄别所起到的对立作用，进而讨论了基于人工智能的技术性方案在长期市场中存在的隐患。最后，为了最大程度地保护市场真实信息资源，本文从机构建设和跨机构合作的角度，针对人工智能的使用提出了可行的设计方案。

二、人工智能技术在信息搜寻市场中呈现的价值对立倾向

虚假信息、虚假新闻一度被一些西方学者认为是观念市场（the marketplace of

ideas）中的一部分，是自由市场中必须付出的代价（Bybee & Jenkins，2019）。他们认为，对于事实最好的检验方式是让其在市场竞争中被承认，消费者的需求偏好可以帮助真实信息脱颖而出；因为按照观念市场的理论，好的观念就像好的产品一样，最终可以在市场中获胜。政府的法律规制、企业和平台的设计变化、消费者保护的议程设置，都违背了自由市场的宗旨。这种从观念市场理论中生发出的有关虚假信息的论断中隐藏着这样的假设：所有人都有获得个体所需信息的权利和途径，信息没有交易成本，人人都有能力甄别和选择。然而，这种观点忽视了人类在搜寻信息能力方面的差别。保罗·布里茨克（Brietzke，1996）认为这种观点忽视了人类社会中偏执、恐慌等文化特质对人们在信息选择时产生的影响。同时，经济或政治强权在很多情形下支撑了虚假信息的生产和传播，因此秉持这种观点也忽视了边缘人群，强化了权力既得者。换言之，真实信息的获取并非没有成本，而需要人们付出成本进行搜寻。

美国著名经济学家、1982年诺贝尔经济学奖得主乔治·斯蒂格勒（George Stigler）提出了信息经济学和搜寻的概念。他认为，在大多数时候信息是一种珍贵的资源，就像经济市场中的其它商品一样，取得信息需要付出成本，也能够获得收益（Stigler et al.，1995）。他以劳动力市场上的信息搜寻活动为例，运用具体模型解释了这一结论：相似工作的工资是有差别的，由于人们都想找到一个工资高的工作，因此也就需要付出更高的搜寻成本。在劳动力市场上，只有当个人的预估边际收益与边际成本相等时，搜寻才会停止（Stigler，1962）。

本文将信息市场和搜寻的概念应用于在网络平台中用户对真实信息的搜寻活动：如同劳动力市场的工作信息一样，真实信息及其搜寻也是有成本的，因此在信息市场中人们不可能无止境地搜寻真实信息，个体搜寻信息的次数和时间只能是有限的。人们搜寻信息的次数与对信息的预估价值和搜寻成本相关；当搜寻的边际收益增多，人们就更倾向于花费更多的搜寻时间；而人们搜寻信息花费的时间越多，搜寻的边际成本也越高。同时，对于不同的消费者来说，搜寻成本也是不同的。

人工智能进入虚假信息的生产环节，极大提高了用户对于真实信息的搜寻成本。尽管自动化新闻同样能够生产真实信息，但相较于虚假信息，两者的比重极不平衡。一个重要的原因是在线信息市场中虚假信息的传播背后拥有巨大的利益动机。大多数虚假信息都有其经济动机（Facebook，2021），虚假信息作为诱饵吸引平台用户点击，为虚假内容的创建者提供以观看次数为衡量标准的收益。这种商业价值和经济利益促使虚假信息的生成和牟利成为一门生意甚至产业，规模化效应使虚假信息的传播更为猛烈。而人工智能技术的入场，给这项产业提供了动力之源，加速了虚假信息的生产，用户对于真实信息的搜寻难度和成本呈几何级增长。超出市场负荷的虚假信息，不仅导致普通搜寻

个体需要付出更多的成本搜寻真实信息，整个社会也需要为此负担更高的搜寻代价：泛滥的虚假信息对信息环境造成污染，使社会信任体系加速崩塌，人们的搜寻环境也变得愈发恶劣。随着时间延续，用户对于真实信息的搜寻成本也将更加高昂。

但也正是由于信息搜寻成本的存在和升高，信息甄别技术得以迅速发展。不断攀升的搜寻成本，对于信息搜寻者产生了两种效果：主动或被动放弃搜寻真实信息（直面泛滥的虚假信息），或离开搜寻市场（拒绝社交媒体平台信息）。显然，这两种结果是社交媒体平台和政府所不期望的，寻找降低搜寻成本的办法以快速解决问题成为了各方共识。在愈发庞大的虚假信息数量背景之下，依赖人工的核查方式效率低下，利用人工智能技术的核查方式成为更佳选择。

在生产与治理的循环中，人工智能对于信息搜寻市场产生了两种对立的影响：原本由于市场收益和经济动机而出现的人工智能“造假”技术，摇身一变成为一种“治假”工具。这种技术性治理方案的运行特征应被概括为：以人工智能和算法为技术底座，以大科技公司为主导方，立足于现有社交媒体平台和权力架构的一场“救火”行动。总结来说，包含以下三方面的具体实践。

一是人工智能被用于精准打击虚假信息源头，试图阻止虚假信息的生产。几乎每一个社交媒体都正在致力于使用和完善机器学习来删除平台上的虚假账号，从账户源头上禁止虚假账号发布信息，并减少从社交媒体向外部虚假信息网站的流量导入，以阻断其经济收益。利用人工智能技术，WhatsApp限制特定消息可以转发的次数，禁止每月批量发送自动化虚假信息超过200万条的账户。[2]谷歌公司则资助了15万英镑给三家英国核查机构以开发自动化事实核查工具，旨在对虚假信息网站进行打击（Jackson，2016）。由于虚假信息给外交和国家主权带来越来越多的挑战，国家机构也开始重视使用人工智能和自然语言技术，抵抗虚假信息破坏性活动。例如，美国国务院的全球参与中心（Global Engagement Center）是一个由国会授权的数据驱动任务中心，专门开发“人工智能工具箱”用以识别、理解和反击旨在破坏或影响美国国家政策和安全的虚假信息活动。[3]

二是当虚假信息进入市场，人工智能技术被用于核查信息以及快速分发同主题的真实信息，以扭转被虚假信息压垮的信息市场。在新冠疫情期间，腾讯较真事实核查平台和中国医师协会健康传播工作委员会联合推出了新冠病毒肺炎实时辟谣，[4]对于热门信息进行集中核查，同时给虚假错误信息贴上标签，当用户尝试发布和分享与此相关的信息时，系统会自动弹出警告信息。此外，各搜索引擎通过各自的大数据和搜索算法，帮助用户获得准确的疫情信息。例如，当用户在百度搜索疫情相关词语，在搜索结果上方会出现信息面板，直观呈现出疫情实时大数据报告；使用谷歌搜索，类似的信息面板还会根据用户搜

索地点呈现出疫苗信息、疫苗位置、剂量统计数据的最新信息以及权威新闻的链接。

三是推出教育工具，提高用户搜寻真实信息的意识和技能。研究表明，数字媒体素养的不足是人们信任虚假信息的重要因素，而简单、可扩展的媒体素养干预可以提高人们识别和感知虚假信息的能力（Guess et al.，2020）。Facebook推出的数字素养图书馆[5]可以被视为一场数字卫生行动的尝试，它将网络社交行为分解为在线参与、隐私保护、内容创作等具体类别，向用户尤其是年轻人传达如何批判性地使用技术和消费信息的理念。在美国，超过10所大学已经推出新闻素养课程，微软与华盛顿大学正在进行关于提高用户媒介素养的合作，普及如何使用技术识别虚假新闻（Burt & Horvitz，2020）；至少有15个州强制要求公立中学进行以数字为重点的媒体素养教育（Rosenwald，2017）。从这个意义上说，人工智能使用户自觉抵御虚假信息，并提高搜寻真实信息的效率。

三、人工智能对抗自动化虚假信息的现实困境

人工智能技术参与虚假信息治理，目的在于减少人们在信息市场中搜寻真实信息的成本，并消解人工智能在生成虚假信息方面所产生的负面影响。然而，数年来的全球实践显示出，人工智能与虚假信息的对抗战果并不如人意。长远来看，利用人工智能降低用户搜寻成本，继而维护真实信息资源的过程，是充满隐患和挑战的。

（一）无法改善真实信息在市场中的相对稀缺性，加重用户的搜寻负担

一方面，人工智能所生产的虚假信息，能够轻易获得广泛传播。近年来人们越来越趋向于使用搜索引擎和社交媒体等来搜寻信息。戈列别斯基（Boyd & Golebiewski，2018）曾在2018年提出“数据空洞”（data void）的概念，用来形容那些搜索引擎中没有结果或者结果很模糊的搜索要求。这些数据空洞极其容易被用于散播虚假信息。人工智能技术和大数据能够让虚假信息迅速占领这些数据空洞，并且利用算法使内容更多地出现在用户的页面中——即在一些还未有足够数据的内容上，制造出全新的假冒内容；同时利用精准定位，让这些虚假信息更精准地出现在人们的社交网站页面或搜索内容中。例如，假新闻制造者可以制造出一些从未出现过的新词或关联短语，再利用虚假账号大量发布有关这些词语的虚假内容。在算法推荐系统的助力下，大量真实用户也搜索该虚假内容。尽管网站和社交网络平台都有发现这些数据空洞的能力并及时删除这些信息，但是往往时间已经滞后，这些虚假信息在人工智能技术的帮助下已经以极短的时间占据了市场。不仅如此，可以被虚假信息乘虚而入的数据空洞还是无穷无尽的。因此，无论是人工智能删除虚假账户还是开展事实核查，都是对已经生产出的虚假信息的

围追堵截。在大多数时候，虚假信息已经造成了一定范围内的影响，利用人工智能治理虚假信息永远处于追随状态。

另一方面，用户处理信息能力也是局限的。泛滥的虚假信息和辟谣信息的双重负担，很容易使用户信息过载。过多的信息使得人的注意力变得有限，人们也许还没有等到人工智能核查虚假信息时，就已经放弃搜寻。由于智能化事实核查信息通常不会提供太多有关虚假信息的背景原因，且缺乏语境的解读，因此对于即使停留在搜寻市场的用户来说，在面对智能化核查信息时仍然会产生不一致且不可控的搜寻决策。总结来说，就人工智能参与虚假信息的生产和甄别的整个过程来说，无论是从信息市场的无限性，还是人类大脑处理信息的局限性，都将导致真实信息变得更加稀缺，而人们搜寻真实信息的负担也更加沉重。

（二）扩大用户间搜寻成本的差异，加深信息环境的鸿沟

用户在技术资源、技术使用能力和使用习惯上存在差异，因此人工智能参与虚假信息的生产和核查使得用户对真实信息的搜寻成本差异性也在变大，长期将带来真实信息资源和信息环境的巨大鸿沟。根据对欧洲居民的一项调查显示，37%的居民每天都会遭遇虚假新闻。[6]人们遭遇虚假信息的次数已经十分频繁，接收虚假信息并辟谣逐渐发展成为获取真实信息的常见流程，甚至成为一种有关信息行为的生活方式：对于没有能力运用技术甄别虚假信息的人群来说，真实信息的搜寻成本使得他们无法享受这种生活方式，长久下去，信息环境也就更加恶劣。反之亦然。尽管各大科技公司及社交网络平台已经将智能化新闻核查技术普及到用户的客户端，方便用户查证，如腾讯公司在在微信平台推出了辟谣助手小程序，用户能在第一时间将阅读过的文章搜索查证，但对于包含复杂要素的虚假信息，利用技术进行查证需要一定门槛。例如，利用换脸技术生成的虚假视频看上去极其真实，在甄别时需要利用基于图像取证、生理信号、寻找图像篡改痕迹等方法来进行检测（李旭嵘，2021），而这无疑是一项高门槛的技术活动。

此外，人们使用技术甄别虚假信息的意愿高低，也会影响到信息搜寻市场的长期发展变化。2016年美国大选期间，马其顿因为其蓬勃发展的虚假新闻产业而成为国际热点，在这个有接近四分之一的人失业的国家，青少年将虚假信息看成摇钱树，对于他们来说，唯一的目的就是使得虚假信息广为传播（Smith & Banic，2016）。因此，这些年轻人学习人工智能技术并建造虚假信息网站，制造“新闻创意”，并且利用“标题党”（clickbait）吸引目标人群点击，从而依靠网站访问量来获得广告收入；与之相反，作为欧洲最能抵御虚假新闻的国家，芬兰政府于2014年就发起了反对虚假新闻的教育倡议，旨在向居民，尤其是学生传授如何学会利用技术对抗虚假新闻，为未来的复杂

数字环境做准备——学生们学习如何识别虚假账号，如何主动求助专家甄别虚假信息（Mackintosh，2019）。人工智能这项技术对于马其顿和芬兰的年轻人来说，意义是截然不同的。对于技术与虚假信息态度的巨大差异，决定了他们在信息市场中的搜寻态度——是主动寻找真实信息，还是直接放弃搜寻，而这些也最终决定他们所处的真实信息环境将走向何处。

（三）利益相关方缺乏一致行动，维护真实信息资源效果有限

信息搜寻成本降低对利益相关方的意义不同，衡量标准也并不统一。这使得各方难以进行有效的协商和妥协，因此经常面临行动不一致的的问题。虚假信息的仲裁者来自三方：社会仲裁者，包括媒体、学者和社会力量；以政府机构为主的执法机关；商业机构（Wiesenfeld et al.，2008）。由于网络虚假信息的治理和政治安全的现实效果密切相连（韩娜，2020），因此政府利用技术打击虚假信息主要投入在保证国家安全。例如，美国联邦调查局FBI投入大量资源和技术力量，以抵抗极端恐怖组织的信息战；而对商业机构和科技公司来说，尽管拥有直接的技术优势，但是由于降低用户的搜寻成本与公司的商业利益是相违背的，大型科技公司解决虚假信息的重点，在于减轻法律责任风险和改善用户体验以维护良好的公共关系。换言之，科技公司利用技术所要营造的信息环境追求的是体验感，而非真实性，因此用户搜寻真实信息的成本下降只是客观结果，并不是公司明确的主观意愿。相对于政府和科技公司来说，作为社会仲裁者的媒体和第三方核查机构，在利用技术对抗虚假信息方面，一直处于弱势。例如，尽管目前大多数社交媒体和搜索引擎平台都与第三方的事实核查机构合作开展信息核查工作，但是合作大多并不顺利。在2019年初，最大的事实检查网站之一Snopes与此前合作的Facebook解除了关系。Snopes认为，虽然潜在的虚假文章或视频通常会传递给它们，但是虚假信息缺乏背景信息，Facebook也没有说明这些虚假信息为何被选中，又是按照何种标准选中，因此事实核查机构不过是在帮助科技公司保持自己社交平台运行顺畅而已。公司和社交媒体平台存在的“道德风险”阻止了它们与第三方核查机构的合作：科技公司所承诺的针对虚假信息和账号的管控措施，由于其内部算法并不公开，因此承诺无法得到验证。

用人工智能构建自动假新闻过滤器面临着许多挑战。从长远来说，人工智能无法理解写作，也无法理解人类文化。即使它能够做到这些，人工智能的使用最终还是会遇到使人类都束手无策的状况。利用人工智能去抵抗虚假信息，不仅是技术本身的抗争，更是相关制度与技术共同作用的结果。将治理虚假信息的重任交给人工智能，事实上掩盖了技术背后的复杂网络和协同合作。当不同的人群和机构不能就哪些信息需要按照何种

标准进行核查而达成共识，那么就更不可能教会算法和技术为人类做出判断。而当共识缺失的时候，人工智能将令信息市场变得更加复杂，用户对真实信息的搜寻负担加重，信息环境的差异也将在不同地区和人群中变大。

四、重返公共性：人工智能管理真实信息资源的方案

人工智能生产的虚假信息是对于真实信息生态的污染。人工智能能否最大限度地降低人们在网络中搜寻真实信息的成本，消解技术对于信息生态的负面影响，取决于如何看待人工智能在维护真实信息生态方面的作用。由于信息的公共性，人工智能生产和核查信息，存在的效应不仅会影响信息搜寻的双方，即社交媒体平台和其用户；更会影响公共物品，如与信息相关的公共事件、公共利益甚至公共精神等。因此，完全基于市场的人工智能使用方案，是无法取得预期效果的。

理查德·斯托曼（Richard Stallman）于2007年提出了与“专有软件”（proprietary software）相对的“自由软件”（free software）概念，并在《自由软件，自由社会》一书中提出，“用户应当有自由地运行、学习并修改、分发软件副本的自由，这些自由在数字化世界中极其重要，以反抗版权对用户自由的侵害”（Stallman，2002：3）。我们将这种思考延续到如今的算法所有权问题的讨论中——人工智能实际由一系列算法组成，算法为几乎所有人工智能系统提供指令并解决问题，是人工智能的核心。讨论人工智能在生产和甄别虚假信息的决定，实际上可以集中理解为算法决策在信息领域的应用。如果以自由软件的标准来讨论算法，那么目前的算法不仅是专有的（即由大科技公司或平台专有），甚至是垄断的。到目前为止，即使算法用于公共目的，公众也无法检查算法的实现或培训数据。几大互联网平台尽管为社会带来了巨大的创新和益处，但是由它们所控制的广泛信息生态使得它们很难被替代，从而形成数字霸主（Kennedy，2020），包括算法在内的专有技术是支撑这些平台得以垄断的关键。最近，越来越多关于希望建立算法规范，防止个人受到不当伤害，并使社会从算法中受益的声音开始出现。

按照此前的信息搜寻市场的模型，对于预估价值高的信息，人们往往愿意付出更多的时间搜寻。因此，从保障社会利益最大化的角度来看，科技公司和社交媒体平台对于与公共利益密切相关或处于重大公共事件时期的信息的推送和核查的相关算法，需要被政府和法律规制。研究证明只要在人群中有1%的欺骗者发布虚假信息，就可以摧毁整个合作性行为（Kopp et al.，2018），这在重大公共事件中会造成极具毁灭性的后果。因此，在与公共利益直接相关或重要时期的算法决策所影响的真实信息环境，极大关乎

整个社会和每个个体的利益。以我国的新冠肺炎疫情时期的网络谣言治理为例，政府主导并与多家网络平台联合建立了一体化的大数据平台，进行谣言数据的汇集、清洗和挖掘，对网络谣言实施智能化甄别评估，实施网络谣言快速处置的数据治理策略（李怀杰，2020）。同时，信息分发和核实的算法在特殊的公共时期也应当尽量透明，接受公众意见并做出合适的调整。例如，在2020美国大选期间，以高度透明和完全大众参与而著称的维基百科平台成立了由用户志愿者组成的虚假信息的特别小组，在大选期间采取一些特别的程序设定，以阻挡有预谋的虚假信息和虚假账号的攻击。

此外，用户参与监督该类算法的可能途径是寻找信息搜寻的代理人，并让其替代个体用户与作为算法专有者的大公司交易。尽管人人参与信息核实是有益的，但要求每个用户在每类信息上都进入搜寻市场，是不现实的。在信息搜寻市场上，当搜寻的边际成本增加，人们既可能停止搜寻，也可能寻找其它的搜寻中介去降低搜寻成本。用户在搜寻真实信息方面的代理人分为两类，一类是现今正越来越活跃的、旨在更快鉴别虚假信息的专业事实核查机构/组织。另一类用户代理人是现阶段在信息核查方面作用被低估却理应生产更多高质量信息的专业新闻记者。随着虚假信息的加速泛滥和逼真程度的提高，新闻记者自身也面临着更重的事实核查负担，亟需资金和技术的支持。科技公司和网络平台应当与媒体合作，开发更加完善的数字取证工具（digital forensics tools）以帮助新闻记者高效地完成信息源的调查和评估，从而为用户提供更多数量和更高质量的真实信息，进而优化真实信息资源的生存环境。2020年，谷歌与旗下的新闻机构Storyful合作创建了一个名为Source的应用程序。依靠人工智能技术，该程序可以帮助新闻媒体分析图像的出处和所有操作，以检测虚假图像；公司还承诺投入3亿美元帮助新闻业（其中一个重要项目为“人工智能与新闻”），从技术和资金上帮助新闻记者使用人工智能来撰写更具相关性和互动性的高质量新闻（Jay，2020）。

经济学术语中的公共产品是一种并不竞争和并不排他的资源。人工智能及算法显然也拥有这样的特质：人工智能和算法的滥用所产生的代价，每一个用户都需要承担；而人工智能对于虚假信息的甄别和核实，也能够让所有用户获利。因此，尽管人工智能的产权目前仍然隶属于大科技公司，但如果完全忽视了人工智能作为公共产品的特性，那么就会令信息资源管理的政策出现偏向，真实信息的生态将进一步恶化。重视人工智能的公共性，实际上就是以最大化公众利益为目标。而在网络信息搜寻市场中，这就具体体现在让个体用更少时间搜索到真实信息，或以最小代价规避可能的虚假信息。因此，一个理想的针对算法决策的合作方案是这样的：与公共利益直接相关的信息的算法决策，应由算法专有者（科技公司和其网络平台）和规则方（政府和法律）共同商定；而围绕一般信息分发和核实的算法，应由算法专有者享有权利，以维护其商业利益，用户

通过其代理人（专业记者或第三方核查机构）在必要情形下对其进行监督。

回到实践中，围绕算法的各方权力博弈，导致该公共性方案必定不可能完全按理想状态得以实施。一些细节问题亟待回答：在信息搜寻中信息边界应该设在何处，即何种信息该被认为具有绝对的公共价值？算法决策权的利益相关方之间的权利如何平衡？作为用户代理人角色的事实核查机构和新闻记者又如何得到切实赋权？本文结合已有的国际实践，提出以下三方面的原则和方向。

首先，继续对人工智能和机器学习技术创新给予支持，但其发展方向应往可解释性人工智能倾斜。2021年1月，全国信息安全标准化技术委员会正式发布《网络安全标准实践指南——人工智能伦理安全风险防范指引》，这份国家层面出台的首个涉及一般性、基础性人工智能伦理问题与安全风险问题的指引文件，对人工智能的研究开发者提出了鼓励性、引导性的目标，而对设计制造者和部署应用者则提出更多的限制性条款（贾开、薛澜，2021）。这种将算法专有者内部的角色予以细分，职责加以细化的理念，有效平衡了技术创新和问责机制。

其次，信息的合理边界和重要算法决策，应通过组建第三方机构提出建议和裁决。第三方机构应从政府、大科技公司、公共社会中推举成员共同组成，需拥有独立透明的运行章程，以服务公共利益为最终目标。其职责包含：有权要求社交媒体平台提供各种信号和数据以确定信息风险层级，对可能显著影响网站推荐流量的算法在更改前应向机构发出警报；有权对公权力过度介入算法审查和干预信息自由作出评估和预警；决定针对特定时期和事件的算法调整方式，以及在特定事件中做出删除内容决定的标准。在国际上我们已经看到类似的努力正在发生。例如，2019年成立的欧洲数字媒体观察站（The European Digital Media Observatory）[7]是欧盟委员会资助成立的一个独立机构，它将事实核查员、学术研究者、媒体机构、在线社交媒体平台以及政策制定者连接在一起，在虚假信息治理政策、信息核查实践、学术研究以及媒介素养倡导等领域进行多方合作。再如被称为“Facebook最高法院”的Facebook监督委员会（the Facebook Oversight Board）[8]于2020年10月正式开始审议，迄今为止作出的最著名决策是2021年5月关于裁定Facebook是否应该让前总统特朗普在被禁言后重回其平台。尽管以上两个案例都并不完美，但却是有益尝试。

最后，在更为日常的虚假信息治理实践中，事实核查机构和新闻记者要跳出辅助社交媒体“灭火”的角色，应更主动地与其形成竞争关系，重塑技术和组织力量，以形成对抗虚假信息的独立免疫屏障。这并不意味着给新闻记者或专业核查员增加更多的技术要求和职业负担（例如要求他们花费更多时间和经费研究算法），而是鼓励他们从现有社交媒体平台（公司）以外的市场中寻求更多合作的可能。可喜的是，打击在线虚假

信息的工作，已经越来越受到学术研究人员、民间社会团体和商业部门的关注，尤其是网络技术公司和初创公司。例如，由志愿者发起的“虚假新闻挑战赛”（Fake News Challenge）[9]旨在通过组织商业竞赛来促进解决虚假信息的技术工具开发，以帮助事实核查员更好地开展工作，老牌科技公司思科公司的网络安全部门Talos Intelligence凭借其基于梯度提升决策树和深度卷积神经网络的集成技术赢得了比赛；再如英国初创商业公司Logically则是在打击虚假信息领域的明星初创公司之一，这家由麻省理工学院和剑桥大学校友于2017年创立的网络公司着手开发的人工智能解决方案，可以高精准地验证新闻、网络对话和图像的真实性（Bernard，2021）。图1简单描述了一个围绕虚假信息治理的跨部门监督和合作方案：集中体现公共利益的信息（如与国家安全直接相关或在公共危机时期的信息），应在第三方机构的监督和协调下，由政府和算法专有者共同商讨信息算法决策；而对于一般类别的虚假信息治理，则由算法专有者和用户代理人，在充分利用商业资源的条件下，寻找最优的算法解决方案。

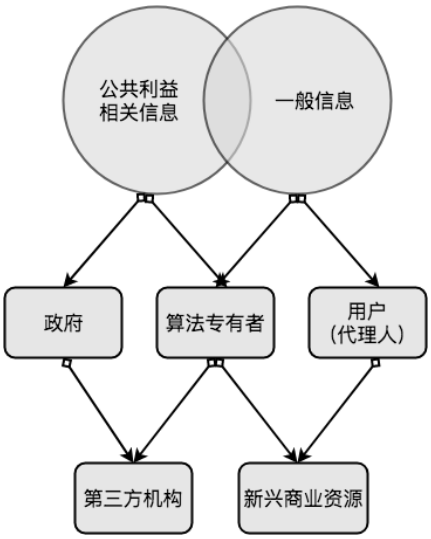

图1 围绕虚假信息治理的跨部门合作方案

五、结语

人工智能自动生成虚假信息，给用户对真实信息的搜寻带来极大的干扰和负担，并快速污染了真实信息的生态环境。但与此同时，人工智能又被赋予重任，作为事实核查技术，试图来解决虚假信息泛滥的问题。从信息搜寻市场和用户搜寻成本的角度，来看待人工智能的这两方面作用和影响，是将信息生产和甄别放置进了同一个场景中予以讨论。这个过程可以更清楚地展现出在这场竞赛中技术参与信息生产和治理的复杂与困

难。这种“找到办法就立即采用，哪怕解决原有冲突的办法可能会扩大和造成新的风险”的行为逻辑实际上掉入了一个循环陷阱中。在帮助事实核查的同时，人工智能参与治理虚假信息有可能带来新的问题：提高了技术“造假”的水平，扩大信息环境质量的鸿沟，掩盖技术背后复杂的关系网等等。

以信息搜寻市场的角度将虚假信息生产与治理连接起来的另一个作用，是人工智能背后的利益相关者网络的凸显。生产虚假信息的人工智能滥用者，治理虚假信息的政府和科技公司，搜寻真实信息的个体用户和第三方事实核查机构等各方的利益、诉求和理念都大相径庭。这种权力困境解释了，以技术手段来治理技术隐患而不触动利益相关方的权利架构的解决方案，是无法达到目标的。互联网平台应更加开放并拥有更透明的问责制流程，并最终建立一个更公平的跨部门组织来监督算法，以对包括虚假信息在内的不当行为进行监督（赵曙光、张竹箐，2019）。而打击虚假信息的社会联盟，实际也就成为支撑真实信息的基础设施（Papacharissi，2020）。在这些基础设施的架构下，人们才能够以相对低廉的搜寻成本获取真实信息。

讨论人工智能如何保护真实信息资源，归根结底是在探讨技术如何影响信息获取，以及算法和信息的权利归属问题。在信息搜寻活动中，决定搜寻过程的算法理应透明，然而，如果开放程度过高，人们有可能使用其过滤器对系统滥用。因此，在信息分发和核实领域，算法的专有和开放程度应取决于信息的不同性质——如在公共利益直接相关的信息处理上，算法应该打开黑箱——以最终得到一个合理的算法规制和合作的方案。美国计算机学会公共政策委员（ACM US Public Policy Council）在2017年提出的有关算法透明度和问责制的七项原则：意识、准入和纠错、问责制、解释、数据来源、可审核性、检验和测试（许向东，2020），可以为未来算法在信息分发和治理的使用提供框架和参考。

随着信息经济的兴起，回到一个没有人工智能和算法参与用户信息搜寻的世界是不可能的。但这种不仅能够制造虚假信息污染，也能够实现虚假信息过滤的技术究竟能走多远，如何走，需要更加开放和持续的讨论。虚假信息的问题已经不只是平台或算法的问题，而是更普遍的国际问题和社会议题。任何解决方案都不可能像运行算法一样简单执行，但一个合理的围绕算法的机构设计和权利架构是必须的。相较于欧美人工智能和信息治理被大科技公司垄断而政府常常束手无策的局面，我国算法决策和信息治理工作则更多由政府规划与建设。如何在维护最大公共利益的目标前提之下，保证技术发展的活力，找到权利制约的平衡点，同时吸引更多的商业力量加入虚假信息的治理，是值得深入思考的问题。

注释：

[1] 数据来源自网站：https://www.statista.com/statistics/1013474/facebook-fake-account-removal-quarter/.
[2] 数据来源参见以下网站报道：http://www.theguardian.com/technology/2019/feb/06/whatsapp-deleting-two-million-accounts-per-month-to-stop-fake-news.
[3] 参见美国国务院全球参与中心（Global Engagement Center）机构网站：https://www.state.gov/bureaus-offices/under-secretary-for-public-diplomacy-and-public-affairs/global-engagement-center/.
[4] 参见腾讯较真新冠病毒肺炎实时辟谣机构网站：https://vp.fact.qq.com/.
[5] 参见Facebook推出的数字素养图书馆网站：https://www.facebook.com/safety/educators/.
[6] 数据来源自网站：https://www.statista.com/statistics/1076568/fake-news-frequency-europe/.
[7] 参见欧洲数字媒体观察站：https://edmo.eu/edmo-at-a-glance/.
[8] 参见Facebook监督委员会网站：https://oversightboard.com/.
[9] 参见虚假新闻挑战赛：http://www.fakenewschallenge.org/.

参考文献：

韩娜(2020).网络虚假信息的生成逻辑与传播模式——从国家安全视角的观察[J].青年记者(33):22-23.
贾开,薛澜(2021).人工智能伦理问题与安全风险治理的全球比较与中国实践[J].公共管理评论(1):122-134.
李怀杰(2020-04-13).运用大数据提升疫情防控网络谣言治理能力[N].光明网理论频道报.
李旭嵘,纪守领,吴春明,刘振广,邓水光,程鹏,孔祥维(2021).深度伪造与检测技术综述[J].软件学报(2):496-518.
倪培昆,朱建明,王国庆(2021).在线社交网络虚假信息交互量最小化的边阻断策略研究[J].中国管理科学(9):188-200.
陈昌凤,师文(2018).智能化新闻核查技术：算法、逻辑与局限[J].新闻大学(6):42-49+148.
王征,叶长安(2019).微博谣言识别与预警算法研究[J].情报杂志(4):148-154.
位志广,宋小康,朱庆华,沈超,张玥(2020).基于随机森林的健康谣言分享意愿研究[J].现代情报(5):78-87.
许向东,王怡溪(2020).智能传播中算法偏见的成因、影响与对策[J].国际新闻界(10):69-85.
杨洸闻,佳媛(2020).微信朋友圈的虚假健康信息纠错：平台、策略与议题之影响研究[J].新闻与传播研究(8):26-43+126.
尹鹏博,潘伟民,彭成,张海军(2020).基于用户特征分析的微博谣言早期检测研究[J].情报杂志(7):81-86.
曾姿颖,黄煜,张引,宋韵雅,周琳(2021).信息诊断系统设计思路：人工核查、广众参与和人工智能的三合一运用[J].全球传媒学刊(1):35-62.
赵曙光,张竹箐(2020).2019年国际新闻传播学研究的十个核心议题[J].新闻记者(7):81-96.
Acker, A., & Donovan, J. (2019). Data Craft: A Theory/Methods Package for Critical Internet Studies. *Information, Communication & Society*, 22(11), 1590-1609.
Ben, B., Andrew, L., Micah, M., & Katerina, S. (2021, May). Truth, Lies, and Automation. *Center for Security and Emerging Technology, Georgetown Univ*. Retrieved June 11, 2021, from https://cset.georgetown.edu/publication/truth-lies-and-automation/.
Boyd, D., & Golebiewski, M. (2018). Data Voids: Where Missing Data Can Easily Be Exploited. *Data & Society; Data & Society Research Institute*, May 11:1-5.
Brietzke, P. H. (1996). How and Why the Marketplace of Ideas Fails. *Valparaiso University Law Review*, 31(3), 951-970.
Burt, T., & Horvitz, E. (2020, 1 September). New Steps to Combat Disinformation—Microsoft On the Issues. *Microsoft Media*. Retrieved June 11, 2021, from https://blogs.microsoft.com/on-the-issues/2020/09/01/disinformation-deepfakes-newsguard-video-authenticator/.
Bybee, K. J., & Jenkins, L. (2019). Free Speech, Free Press, and Fake News: What If the Marketplace of Ideas isn't About Identifying Truth? SSRN Scholarly Paper ID 3400882. *Social Science Research Network*,18.
Facebook Newsroom. (2018). Seeing the Truth. About Facebook. *Facebook Media*. Retrieved June 11, 2021, from https://about.fb.com/news/2018/09/inside-feed-tessa-lyons-photos-videos/.
Facebook Newsroom. (2017, 7 April). Working to Stop Misinformation and False News. *Facebook Media*. Retrieved June 11, 2021, from https://www.facebook.com/formedia/blog/working-to-stop-misinformation-and-false-news.
Ghebreyesus, T. A. (2020, 8 February). Director-General's Remarks at the Media Briefing on 2019 Novel Coronavirus. Retrieved June 11, 2021, from https://www.who.int/director-general/speeches/detail/director-general-s-remarks-at-the-media-briefing-on-2019-novel-coronavirus---8-february-2020.
Guess, A. M., Lerner, M., Lyons, B., Montgomery, J. M., Nyhan, B., Reifler, J., & Sircar, N. (2020). A Digital Media Literacy Intervention Increases Discernment between Mainstream and False News in the United States and India. *Proceedings of the National Academy of Sciences*, 117(27), 15536-15545.
Jasper, J. (2016, 17 Nov.). Fake News Clampdown: Google Gives €150,000 to Fact-checking Projects. *The Guardian*. Retrieved June 11, 2021, from http://www.theguardian.com/media/2016/nov/17/fake-news-google-funding-fact-checking-us-election.

Kennedy, J. (2020). Monopoly Myths: Do Internet Platforms Threaten Competition? *Information Technology and Innovation Foundation*, 2020 Annual Report.

Kopp, C., Korb, K. B., & Mills, B. I. (2018). Information-theoretic Models of Deception: Modelling Cooperation and Diffusion in Populations Exposed to Fake News. *PLOS ONE*, 13(11), e0207383.

Liu, I. J. (2020, 25 February). The New Tool Helping Asian Newsrooms Detect Fake Images. *Google Media Report*. Retrieved June 11, 2021, from https://blog.google/around-the-globe/google-asia/new-tool-helping-asian-newsrooms-detect-fake-images.

Mackintosh, E. (2019). Finland is Winning the War on Fake News. Other Nations want the Blueprint. *CNN Special Report*. Retrieved June 11, 2021, from https://www.cnn.com/interactive/2019/05/europe/finland-fake-news-intl.

Marr, B. (2021, 25 January). Fake News Is Rampant, Here Is How Artificial Intelligence Can Help. *Forbes*. Retrieved June 11, 2021, from https://www.forbes.com/sites/bernardmarr/2021/01/25/fake-news-is-rampant-here-is-how-artificial-intelligence-can-help/.

Michael, R. (2017). Making Media Literacy Great Again. *Columbia Journalism Review*, Fall.

Papacharissi, Z. (2020). The Year We Rebuild the Infrastructure of Truth. *Nieman*, 2020 Annual Report.

Smith, A., & Banic, V. (2016, 9 December). How Macedonian Teens Earn and Spend Thousands from Fake News. *NBC News*. Retrieved June 11, 2021, from https://www.nbcnews.com/news/world/fake-news-how-partying-macedonian-teen-earns-thousands-publishing-lies-n692451.

Stallman, R. (2002). *Free Software, Free Society: Selected Essays of Richard M. Stallman*. Boston, MA: Free Software Foundation.

Stigler, G. J. (1962). Information in the Labor Market. *Journal of Political Economy* 70.5(2), 94-105.

Stigler, G. J., Stigler, S. M., & Friedland, C. (1995). The Journals of Economics. *Journal of Political Economy*, 103(2), 331-359.

Tristan, G. (2018, 27 November). Here's why Low-credibility News Seems to Dominate Twitter. *TNW The Next Web*. Retrieved June 11, 2021, from https://thenextweb.com/artificial-intelligence/2018/11/27/heres-why-low-credibility-news-seems-to-dominate-twitter/.

Warner, M. R.. (2018, 30 July). Potential Policy Proposals for Regulation of Social Media and Technology Firms. *White Paper - U.S. Congress, Senate*. Retrieved June 11, 2021, from https://regmedia.co.uk/2018/07/30/warner_social_media_proposal.pdf.

Wiesenfeld, B. M., Wurthmann, K. A., & Hambrick, D. C. (2008). The Stigmatization and Devaluation of Elites Associated with Corporate Failures: A Process Model. *Academy of Management Review*, 33(1), 231-251.

Yao, Y., Viswanath, B., Cryan, J., Zheng, H., & Zhao, B. Y. (2017). Automated Crowdturfing Attacks and Defenses in Online Review Systems. *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 1143-1158.

imagination" constitute the two narrative structures of the memory of Republic of China in the "*Nanjing Daily*" in the 20 years since the new century. The former is an ideological official narrative, while the latter is an aesthetic historical narrative. Interaction of structural forces such as state, market and culture provides multi-discourse spaces for the news media to produce memories of Republic of China, and finally generated the two narrative structures mentioned above.

[Keywords] Republic of China; collective memory; news production; new institutionalism

61 Value Conflict and Governance Solution: Examining the Role of Artificial Intelligence in Tackling Misinformation

· ***GUAN Bi-lu, GU Li-ping***

[Abstract] Artificial Intelligence (AI) has learnt to generate fake media creations overnight, while AI-based tools could also be used to debunk fake news. Considering the rampant fake accounts and social bots on social websites, individual uses have to pay costs to achieve the real stories. We lay out Stigler's model of search as a framework and illustrate how AI influences the information market and the environment that encouraging real news. Although more AI can help to combat fake news and disinformation more effectively, the burden on individuals for searching real news would not be reduced and the disparity in the information environments among various populations increases in a long run. We then develop the strategies and a cross-agency cooperation model for AI use in fighting fake news and disinformation. We highly recommend that when dealing with public-interest-related news, AI algorithm should be transparent and the institutions should be held responsible for algorithms. Practically, a third-party organization that is independent from both the industry and government should be established, and also, the fact-checking organizations and professional journalists should be given more support.

[Keywords] Artificial Intelligence; fake news; disinformation; fact check; info-search cost

76 Effects of Family Communication on Maternal Health from the Perspective of Uncertainty Management

· ***CHEN Juan, LUO Ke-xin, LI Jin-xu***

[Abstract] Family communication is the main factor affecting the emotional health of