生成式人工智能虚假信息的舆论生态挑战与治理进路

张文祥

摘要：生成式人工智能虚假信息对舆论生态和话语秩序构成了严峻挑战，割裂着舆论事实的认知图景、操纵着舆论话语的公共导向、绑架着舆论价值的社会信任。在群体武器、时空脱域、数据围猎的三重作用下，舆论生态出现异化，呈现出舆论载体“武装化”、舆论主体“脱域化”、舆论客体“失控化”的特征。生成式人工智能虚假信息对舆论生成和走向的设置会分割舆论主流观点、控制舆论传播路径、诱发新型舆论风险，主要表现为公权力受到遵循资本与技术逻辑运作的私权力挟持，定制化的情感偏向信息营造虚假意见环境，虚假视觉信息生成引发舆情并误导舆论走向。生成式人工智能虚假信息的舆论治理应以数据基础设施的统合实现“中台”治理，发挥多元主体的侧芽效应以破解顶端优势带来的治理困局，构建以“FAITH”为核心要素的治理体系以消除新型舆论治理的悬浮状态，重构生成式人工智能下的舆论结构和社会实践。

关键词：生成式人工智能；虚假信息；舆论生态；智能传播；舆论治理

DOI：10.19836/j.cnki.37-1100/c.2025.01.013

随着生成式人工智能(generative artificial intelligence，缩略为GAI)对人类传播格局的深度嵌入，技术已升级为行动主体，同人类发挥着主体间性的关系映射。作为颠覆性技术媒介的生成式人工智能，以其强大的信息资源分配能力和数字权力，进一步验证了施蒂格·夏瓦(Stig Hjarvard)的“媒介化”(mediatization)理论，即媒介将自身建构为一种半独立(semi-independent)的制度，介入不同的社会机制和文化现象，引发制度语境的结构性改变。在此过程中，媒介与政治、经济体系深度交融，成为共同影响社会的关键力量。媒介逐渐建构起独特的物质和符号资源的分配方式以及游移于规则内外的运作模式，形成契合制度主义传统范式的“媒介逻辑”(media logic)①。作为新型智能媒介的生成式人工智能，使人类社会结构面临新的“媒介化”，因其自身独特的技术逻辑和运算性质形成独有的媒介逻辑，本文将其指称为“GAI逻辑”。这种逻辑以特有的数据流形式编织信息形构的世界图景，通过算法与算力的架构运作来操作媒介社会的可见性差异，以造成社会认知、价值与意义的“GAI偏向”，对人类的思维与信息秩序产生重置式影响。生成式人工智能因与政治、经济、文化高度交融互嵌，不可避免地受到人类意图的操纵和影响，从而生发“真实遮蔽”现象。海德格尔认为，现代技术的本质是“促逼”(herausfordern)，居于“座架”(gestell)之中，而这种促逼使得人被“摆置”(stellen)②。作为“座架”的生成式人工智能摆置着现有信息秩序，调动并更迭了虚假信息的生产与分发机制，重新对虚假信息的生产主体赋权，对虚假信息的传播效果赋能，将断裂性的真实嵌入到舆论演化的意义空间，促成虚假信息对舆论生态的扰动。在生成式人工智能下的虚假信息生产、分发、消费过程中，真实人类

山东大学文化传播学院新闻传播学研究所研究助理沈天健对本文有实质性贡献，特此致谢。

基金项目：国家社科基金重大项目“互联网环境下新闻理论范式创新研究”(21&ZD318)、“人工智能时代的新闻伦理与法规”(18ZDA308)。

作者简介：张文祥，浙江大学网络空间安全学院、浙大宁波理工学院传媒与法学院双聘教授，博士生导师，浙江大学网络空间国际治理研究基地秘书长(杭州 310058；zhangwenxiang@zju.edu.cn)。

① 施蒂格·夏瓦：《文化与社会的媒介化》，刘君等译，上海：复旦大学出版社，2018年，第4—22页。

② 《海德格尔选集》，孙周兴译，上海：上海三联书店，1996年，第933—935页。

主体的功能让渡于生成式人工智能技术。因此,重新确认生成式人工智能虚假信息下的舆论生态变革与治理,是21世纪20年代智能传播研究的题中应有之义。

一、生成式人工智能虚假信息下的舆论异化

生成式人工智能虚假信息的"GAI逻辑"对现有舆论生态和话语秩序构成挑战,割裂着舆论事实的认知图景、操纵着舆论话语的公共导向、绑架着舆论价值的社会信任。

1.舆论载体"武装化":作为"群体武器"的虚假信息隐性操纵

以往,作为舆论载体的信息具有较高的生产和分发门槛,用户所生产的虚假信息难以实现专业化和广域化,往往只能作为微型流言在人际网络中传播扩散。在前生成式人工智能传播环境下,虚假信息的制造和分发权限集中于传媒机构和产业公司,它们能够通过后台掌握的用户偏好数据和议程设置焦点来生产精准化的虚假信息,通过大众传播的信息渠道实现规模化的真相遮蔽和情绪感染,对舆论走向和公众话语造成影响。而随着生成式人工智能在互联网场域的接入与下沉,普通用户在理论上拥有了设置和操纵舆论的能力。生成式人工智能可被视为一种"超简易模式"的信息生产工具,用户只需通过提示和引导就能运用生成式人工智能制造缜密的虚假信息。在数据喂养和算法训练中,生成式人工智能逐渐武装起强大的内容逻辑桥接能力和信息真实模拟能力。这种能力看似为用户所用,实则为背后的资本方和技术方所有,用户所能够运作的权限亦被资本逻辑和技术权力隐性地规定与限制。因自身影响力的局限和网络圈层的壁垒,普通用户生成的虚假信息难以转变为自主引发负面舆论的"武器"。用户对生成式人工智能输出虚假信息、干扰舆论生态能力的掌握,实则是在接受资本方和技术方所递送的"武器",用户生成的虚假信息在模型之内便已受到隐性操纵。生成式人工智能虚假信息的生成逻辑更像是在特定舆论情境下对用户的"武装",用户并不享有对生成式人工智能虚假信息的支配权和舆论引导权,而是在重要舆情事件发生时被动获取信息舆论"武器"来完成自我的"武装化",进而作为社交主体生产虚假信息并流散至网络空间。

在情绪屡屡跑赢真相的后真相时代,生成式人工智能背后的控制方会因自身立场、资本逻辑和政治需求等因素在舆情事件中作出操纵舆论的行为。单一依靠算法个性推送、信息茧房编织等方式已难以驯服用户的认知与思维,所以拥有信息资源和数字权力的"数字强势方"会依托人际传播和用户节点实现扁平化的信息输出①。一方面,生成式人工智能得以在用户的提问与对话中生成相应的专业化智能虚假信息,并通过本就有传播意图的用户扩散至人际网络,将用户"武装"为实施虚假信息攻击的"兵士",搭载用户的情感关系和社交资本以操纵舆论走向;另一方面,用户对生成式人工智能虚假语料的投喂,将在机器后台的运算下被吐纳至与其他用户的对话界面之中,间接实现个性化和私人化的虚假信息传播,潜隐不彰地助推舆论场的混乱。"数字强势方"在特定情境下以生成式人工智能方式对用户群体进行"武装",使智能虚假信息逐渐在多元主体驱动下成为激化行业矛盾、引发国际竞争、扰乱社会秩序的"群体武器",最终冲击社会的信息秩序与认知结构②。因此,生成式人工智能虚假信息并非作为一种具有"高杀伤力"的武器来改变公众的政治观点,而只需作为一种规模化的"武装"来误导公众认知的应然与实然,即可使公众在虚假信息弥散的意见环境中作出具有价值偏向的误导行为。

① 张文祥、沈天健、孙熙遥:《从失序到再序:生成式人工智能下的信息秩序变局与治理》,《新闻界》2023年第10期。

② 例如在政治选举中,公众更倾向于根据直觉、价值观和信仰来决定自己的选票。2023年9月,斯洛伐克议会候选人Michal Šimečka谈论购买选票和提高啤酒价格计划的虚假音频信息在选举前夕被发布,作为一种武装化的政治舆论对当地政治生态造成冲击。参见David Adam,"Misinformation Might Sway Elections——But Not In The Way That You Think",https://www.nature.com/articles/d41586-024-01696-z,访问日期:2024年6月28日。

2.舆论主体“脱域化”:在地性舆论的无界弥散与虚假信息的摆置歪曲

社会学家安东尼·吉登斯(Anthony Giddens)曾以“时空脱域”(disembedding)意指社会关系从局部互动的情境关联中脱离出来,在不确定的时间与空间的跨度上重新构建①。在“时空脱域”下,社会关系能够突破时空限制,在任何时间或地点生发和维系。全球网络的互联重塑着传播时空,信息的流散不再受时间限制与地域禁锢,能够以瞬时化和无界化的传播形态将全球范围内的社会关系拉入其中,与“时空脱域”的理论形成互通。

在生成式人工智能勃兴之前,虚假信息的生产与传播实际上并未完全实现全球互联的脱域,承载虚假信息的平台依旧是在地化的,传播基础设施仍有国家和地域的界分。即使全球信息场域已出现超大型社交媒体平台,不同地域的受众仍遵从舆论参与的“在地化”价值逻辑,对脱离圈层属性和周边属性的虚假信息倾向于围观而不是参与,虚假信息传播下舆论生成与扩散的媒介时间整体上仍顺从于特定地域的社会时间。

生成式人工智能的全球应用使其成为超越国界的“世界信息基础设施”,不同地域和背景的虚假信息得以跨过在地化的信息平台,通过全球共通的对话界面被呈现于不同的国别与社群之中,有关在地性舆论的信息弥散至全球。舆论主体以往是较为单一的受众,但生成式人工智能的强势介入使其升维为强大的技术行动主体,并重新构建起舆论场域中的主体脱域关系。在生成式人工智能介入行动实践之前,作为舆论主体的受众对大部分非本土的舆论事件处于“偶遇”状态,即在寻找或浏览某些信息时意外发现舆论事件,具有是否参与舆论的主动性。而生成式人工智能能够将非本土舆论事件穿插于用户对话之中,使用户被动参与到舆论事件中,凭借对用户兴趣偏好的感知,潜移默化地介入全球受众的认知系统,通过持续的虚假信息输出改变用户对某一舆论事实的看法。在全球新闻高度互联之时,生成式人工智能得以将在地性虚假信息“全球化”。当各国用户群体在生成式人工智能大模型的摆置中形成了对某一在地舆论的统一认识后,“全球化”的智能虚假信息又将重归“在地性舆论”,借助全球力量对本土舆论实施歪曲与误读。舆论主体在技术升维与信息弥散的过程中脱域,在舆论场域中接入全球时空,在关系脱域中使舆论“无理化”与“无界化”。

3.舆论客体“失控化”:“数据围猎”下舆论设置逻辑的颠覆

生成式人工智能的出现,被学者称之为人类的“新普罗米修斯时刻”。弗里德曼认为生成式人工智能将技术系统增强关联并形成超级循环,使媒介生态向人机共生互构的主体间性图景转变②。舆论客体指涉舆论生发的事件、现象或人物,是舆论激发的源头。在前生成式人工智能传播环境下,舆论客体的内容传播和分发主要依靠算法完成,注入舆论客体的虚假信息受企业和平台所控制的“算法黑箱”影响。舆论客体议程的操纵实践与发展过程不向公众透明,通过人为的算法操纵来控制意见的可见范围和可见程度,甚至对群体发声的伪装也可通过算法的微调来实现。虚假信息分发逻辑受制于平台逻辑与资本逻辑,对舆论客体与议程具有相对稳定的操纵秩序和设置程式。当生成式人工智能信息秩序产生颠覆性影响,对信息可见性和传播效能进行操控的主体首次同背后的资本方和技术方实现相对脱离。生成式人工智能对话界面的开放为受众声量的上溯提供了路径,虽然用户不可避免地出现“武装化”倾向,但其依然在数据和算法之间具备一定的能动性,能够以“弱者的武器”在信息自主生产与技术逻辑支配的边界之间找到空间。用户可将自身所掌握的信息置入生成式人工智能大模型,在数据吐纳与运算更新的过程中具有一定的自主裁决权。

可以用“数据围猎”概念来刻画生成式人工智能时代用户所具有的强效能舆论设置行为。在舆论生发后,用户能够根据所处圈层和观点的趋同而组成类似“协作捕猎式”的情感共同体,以生成式人工

① 安东尼·吉登斯:《现代性的后果》,田禾译,南京:译林出版社,2011年,第18页。

② Friedmant, “Our New Promethean Moment”, https://www.nytimes.com/2023/03/21/opinion/ artificial-intelligence-chatgpt.html, 访问日期:2024年6月28日。

智能的对话界面为“猎场”,通过在对话界面置入大量偏好数据来扭转生成式人工智能大模型的运算感知,改变对该舆论事件的信息输出方向与意见。这种自下而上的数据控制必须有成建制、成规模、成体系的数据输出行动才能产生实质性影响,实现对大模型本身运行逻辑和背后技术方意图嵌入的调整,扭转生成式人工智能在公共空间的信息呈现,完成对舆论设置议程和逻辑的“围猎”。由此,用户能够自行生成虚假信息并借助生成式人工智能接入平台的代理技术实现传播,搭载算法而实现对舆论生成和走向的设置。在此过程中,生成式人工智能自身的运算逻辑会受到干扰,甚至有可能产生更为严重的“幻觉”(hallucination),进而以自身强大的数据集合、算力基础与算法逻辑脱离人类主体的操控,彻底颠覆舆论设置逻辑。生成式人工智能会遵循自身的数据池和运算逻辑,以接入的社交机器人等充当高度拟人化的意见领袖,联合人类主体共同制造话题、设置议程、调控情感,以技术威权制造沉默的螺旋,分割舆论主流观点并控制舆论传播路径。

二、生成式人工智能虚假信息诱发的舆论风险

生成式人工智能虚假信息对舆论生成和走向的设置会诱发新型舆论风险,主要表现为公权力受到遵循资本与技术逻辑运作的私权力挟持、情感偏向的信息营造虚假意见环境、虚假视觉信息引发舆情并误导舆论走向。

1.私权压制公权:舆论场域受资本和技术逻辑的操纵与挟持

当互联网成为社会基础设施,政府对信息秩序和舆论生态的控制力不足,网络信息内容治理需借助平台力量,顺应平台技术逻辑,由此形成政府—平台—网络用户的三元规制结构①。当公权力在网络平台面前出现失灵现象时,掌控平台的大型科技公司的私权力触角得以扩张、延伸,通过把控各类数据流的“关键出入口”构筑起完整的网络生态系统,拥有网络空间在场与裁决的权力,囊括了“私人”与“公共”财产的混合领域,凭借对规则制定、资源调配的实质性把控,逐渐形成虚拟空间的“私治理”模式②。掌控生成式人工智能的大型科技公司能够调控其数据语料、算力基础和算法机制,凭借“技术霸权”向特定用户、圈层和舆论场输出特定的信息并操控其可见性和引导性。在多元行动者趋向“去中心化”话语表达的技术赋权过程中,舆论场域中以二律背反的动势重新组织起资源与权力的“再中心化”图景,推动虚假信息在舆论生态中的建制化和有序化流散,进而影响舆论生态的真实意义。

在平台私权与“GAI逻辑”崛起之时,公权力的退场可能导致舆论场域受到资本和技术的胁迫,虚假信息在舆论场中的传播机制也受到平台逻辑和资本逻辑的影响。实际上,数字基础设施难以保持中立,它们凭借数据汲取能力、资源调配能力和规则制定能力,在数字社会的分工结构中日益占据优势地位。大型科技公司可凭借其自身的数据垄断优势地位,连接大量入驻平台的第三方经营者,在舆论场中形成虚假信息传播的“寡头垄断”③,然后借助生成式人工智能的高精度、高效率信息生产能力,向社会和个体投放隐含特定意图的虚假信息,从而达到吸纳注意力资源的目的。在此基础上,根据自身或第三方利益来推动集合行为的产生,引发舆论场域的混乱。数据基础设施和技术集权的隐性偏向将助推虚假信息在舆论场的泛滥,致使用户的认知和行为受到误导,使公众逐渐习惯一个没有确定性事实的生存方式,乃至丧失对社会信息环境的信心与信任④。

生成式人工智能虚假信息有时也被应用于企业间的竞争。掌握生成式人工智能技术的企业,会

① 张文祥、杨林、陈力双:《网络信息内容规制的平台权责边界》,《新闻记者》2023年第6期。

② 史安斌、俞雅芸:《全球数智之治的发展态势与制度逻辑——基于2023年网络治理研究的主题聚类分析》,《传媒观察》2024年第1期。

③ 许多奇:《论网络平台数据治理的私权逻辑与公权干预》,《人民论坛·学术前沿》2021年第21期。

④ 黄河:《网络谣言的智能化演变及治理》,《人民论坛》2023年第4期。

借助生成式人工智能炮制竞争对手的虚假信息，对竞争对手实施大规模的“污名化”信息攻势，通过调动受众情感和激化信息态势的方式达到打击竞争对手的目的，客观上使舆论场域失焦。生成式人工智能会搭载部分带有价值倾向和经济属性的虚假信息再返回舆论场域进行植入，借助公众对主体舆论事件的注意力聚合，将带有认知偏向的虚假信息搭载至主体舆论，以“新闻搭车”的方式冲击舆论场域。在此过程中，公权力所分发的公共信息和舆论治理措施会被淹没于生成式人工智能的信息矩阵和数据洪流中，在需要真相与事实的舆论焦点内潜隐不彰，弱化甚至消解了其对主流舆论的引导力。

2.精准的情感流变：以情感摆置营造虚假意见环境

情感是催生舆论并推动舆论扩散的重要力量，其对舆论的影响力甚至已超过事实自身[①]。情感认知成为公众阐释舆论事实、控制意见表达的重要依托，评价性信息构成了情感的意向性成分，情感对象被情感主体视为有价值观倾向的或具有重要性的认知对象[②]。消解情感在舆论生成中的流动变化，表明情感并非舆论中的确定性事物。实际上，群际情感和人际情感的流动变化成为舆论生发演进的重要推动力量。

生成式人工智能技术作为基础性的“座架”被嵌入社会信息结构，在对用户数据进行分析与画像建构的过程中，能系统了解用户的价值偏好和情感取向，在与用户以“对话”方式进行互动时，一定程度上可以规避传统算法和机器人的单向信息传播，并结合用户自身属性体系化地输出符合用户认知、顺从用户情感的虚假信息。生成式人工智能根据用户的情感画像和情绪需求生产定制化的信息内容，会将用户困于自身的情绪茧房，实现对用户情感的摆置，使虚假信息内含的特定意图得以借助用户的情感倾向植入特定的舆论生态。该过程通过把特定情感线索植入虚假信息文本而实现，推动用户的信息接受行为与情感话语深度融合，导致规模化的情感极化[③]，使舆论生态在情绪化虚假信息的刺激下出现偏向与错置。

用户在舆论中的情感倾向经常成为资本捕获与剥削的目标，公众在舆论场域的公共表达意义被技术操控下的情绪化表达所消解，使公共讨论陷入没有交流价值的舆论环境之中[④]。用户的情感受到生成式人工智能的摆置，间接成为生成式人工智能及其背后主体的“传声筒”，成为搅动舆论秩序的“帮凶”。生成式人工智能所持有的海量数据库能够精准识别和定位用户的情感特征，生产出具有煽动性和诱导性的智能虚假信息来诱导用户在特定舆论事件中的情感流动变化，形成独特的“GAI情感流变”逻辑。由于网络舆论中所带有并扩散的情感是公众建构群体身份认同的重要方式，具有相似情感倾向的用户会因为短暂的情绪呼应和标签互认形成临时的交往群体并攻击与自身情感倾向不同的用户。这种群体化的情感确认与冲突行为进一步增加了数字时代网络舆论生态的复杂性[⑤]，并在生成式人工智能的信息生产助推下干扰舆论的事实导向和理性价值。用户往往倾向于相信生成式人工智能所生产的虚假信息以契合自身的情感属性和社交资本，从而导致生成的舆论愈发趋于非理性而成为偏颇的舆论。

面对具有较高媒介素养的受众时，生成式人工智能虚假信息的传播尽管会被阻滞，使常规性的“GAI情感流变”逻辑无法撼动用户的认知或调动用户行为，但生成式人工智能却可借助社交机器人进入公共对话空间，通过自身大模型与交互微调的对话修正功能使社交机器人的言论实现定制化和人性化；凭借自身语言大模型的强势属性使社交机器人的言论劝服和煽动效果更加突出；以信息生产的超高效能在短时间内介入舆论环境，营造虚假的意见环境，干扰用户的正常认知，刺激用户害怕被

① 袁光锋：《迈向“实践”的理论路径：理解公共舆论中的情感表达》，《国际新闻界》2021年第6期。

② 蒋琳：《舆论传播的情感驱动：理论机制与影响效应》，《社会科学家》2024年第3期。

③ 张涛甫：《人工智能推动舆论生态转型及其治理进路》，《学术月刊》2024年第2期。

④ 张志安、冉桢：《互联网平台与情感研究：理论路径与本土框架》，《新闻大学》2022年第12期。

⑤ 张涛甫：《人工智能推动舆论生态转型及其治理进路》，《学术月刊》2024年第2期。

群体孤立的心理弱点,而顺应圈层内的情感流动趋向,在主动或被动的情感支配下实现生成式人工智能虚假信息在舆论场的传播与影响。

3.“真实模拟权限”下沉:视觉霸权误导舆论走向

从ChatGPT到Sora,信息交流从“以文本为基础的单模态升维到以影像为基础的多模态”。Sora以视觉模态为基础,能够将语言对话转为视觉传达,从抽象思维转为具象认知,使人们能够置身近乎真实的场景中进行交流①。Sora使公众参与社会运行的方式从话语表达转为场景模拟,用户的每一种思绪都能在生成式人工智能模拟物理世界的技术加持下实现场景构建的拟合②,人类的认知方式在生成式人工智能的多模态信息生产和传播中被重构,生成式人工智能对舆论的影响将进入视觉场景的真实模拟时代。生成式人工智能通过使信息接收者“致幻”而塑造认知、操纵舆论的能力愈发强大,对社会信任的侵蚀程度日益加深③。生成式人工智能后台的技术复杂成就了用户前台的生产简化,Sora的出现使得高精度高逼真的视觉内容生产赋权至普通用户,每一个用户都可以按照自我意图去“模拟真实”。情感偏向的视觉信息更容易助推舆论的非理性走向,能够以特定视角将信息融入趋近真实的模拟展示之中。当视觉信息对真实的模拟形成组织能力时,会产生舆论生态的视觉霸权现象,虚假信息可在虚假的视觉生成中调动公众的注意力,引发舆情并误导舆论走向。当一个舆情事件引发观点撕裂和意见冲突时,用户可依照自身的立场和观点对事件现场进行“虚假的还原”,将虚假的观点和事实隐藏于多模态的视觉符码之中,规避事实核查并增加辨识难度,将隐含有虚假信息的视觉内容植入舆论意义空间,引发舆论热点的碎裂与失焦。

作为一种非语言传播形态,视觉传播具有感性化、直观化、浅层化等特点,能够通过感受性的、反理性的、反逻辑的场景体验,超越国界与语言的障碍,架构于世界通用的沟通符号体系之中,形成全球共通的信息传播机制④。视觉图像已成为舆论“反转”和“定调”的关键性媒介。作为人类情感的符号化创造,它能够在网络舆论中建构起一种公共意象,在文化背景、图像文本和基本生理特征三个维度的互动过程中建构一致性。生成式人工智能的技术赋权使用户能够自主生产具备意象一致性的视觉文本,在舆论场域中生产虚假视频和新黄色新闻,干扰舆论演化的正常逻辑。舆论形成的最大动因是“虚构”,而不是“客观事实”⑤。Sora等生成式人工智能已在舆论传播中建构起“拟态事实”(pseudo-fact),在一定程度上决定着人类对外部现象世界的认知,形塑着人类自我的“脑中影像”,诱导人类植根于本质的虚构经验(experience)创造虚构的事实⑥。正如艾伦·凯(Alan Kay)将所有的新闻都视为“一种模拟”,认为新闻报道不一定是真实的,而可能是对真实的模拟⑦。生成式人工智能使得视觉化的“模拟事实”逐渐影响人类对舆论事实情境和真实图景的把握,导致舆论愈发呈现出情绪化、不稳定和去中心的噪声形态。

视觉效果在虚假信息中发挥的核心作用来自其索引性(indexicality)。在以ChatGPT为代表的文本内容生成技术中,文字与其指代对象之间并不存在物理相似性的抽象符号,而Sora能够通过将文本指令转化为视频的方式,利用视觉效果对物理对象与实践进行直接描述,即提供了现实的“索

① 喻国明、苏健威:《从Sora到GAI:智能媒介的升维与全新场景体验时代的到来》,《编辑之友》2024年第6期。

② 喻国明、苏芳:《作为真实世界模拟器的媒介与后真相时代的“拨乱反正”——以Sora为例解析数字文明时代的媒介新范式》,《新疆师范大学学报(哲学社会科学版)》2024年第4期。

③ Bontcheva K., Papadopoulous S., Tsalakanidou F., et al.,“Generative AI and Disinformation: Recent Advances, Challenges, and Opportunities”, *European Digital Media Observatory*, 2024, pp. 1-36.

④ 刘庆、何飞:《网络舆论中图像的情感动员机制研究》,《西南民族大学学报(人文社会科学版)》2021年第11期。

⑤ Lippmann W., *Public Opinion*, London: Routledge, 2017, pp. 18-52.

⑥ O'Donnell E、刘学蔚:《否思“后真相”:基于李普曼舆论学视角》,《新闻与传播评论》2020年第3期。

⑦ Kay A., Goldberg A., “Personal Dynamic Media”, *Journal for the Theory of Social Behaviour*, 1977, (10), pp. 31-41.

引”[①]。这种索引内嵌于不同用户的文化原型之中，从而实现情感的共享和价值的认同，以视觉性的情感图式修辞方式召唤社会文化体系中的原型与言说[②]，以全民模拟真实的方式将舆论的引导权限下沉和弥散，再由生成式人工智能自身的技术逻辑进行集中配置，最终输出建制化、体系化的虚假视觉信息来实现对情感的操纵和对舆论的摆置。

三、生成式人工智能虚假信息的舆论治理进路

生成式人工智能虚假信息所诱发的舆论风险不容小觑，应认真对待并予以科学合理的治理。生成式人工智能虚假信息的舆论治理应以数据基础设施的统合共创实现“中台”治理，为舆论生态提供良性的信息图景与事实视野；发挥多元主体的舆论治理侧芽效应以破解顶端优势带来的治理困局，构建多元主体的对话空间；构建涵括向善、真实、惯习、透明、人本等要素的“FAITH”核心要素体系，实现生成式人工智能下舆论治理的要素下沉与整体在场。

1.“中台”治理：数据基础设施的统合共创

学界对智能传播视域下的人工智能治理主要聚焦于“前台”（人机交互接触的界面与软件）与“后台”（支撑人机交往与运算逻辑的硬件）的施治，对协调“前台”与“后台”有序运转和技术缓冲的“中台”（middle platform）治理还缺乏讨论。事实上，生成式人工智能的“中台”已全面嵌入人们的日常生活与实践，并内置有隐匿的技术功能和海量数据，将生成式人工智能转设为一个可编程（programmability）的计算系统。从某种意义上说，“中台”不仅保留着普遍性、可靠性等传统基础设施的特点，更对数据具有持续、密集、集中的提取和再利用能力，已加速完成基础设施化转向（infrastructural turn）[③]。

明确生成式人工智能“中台”的治理进路，以消弭智能虚假信息在舆论生成演进中的危害。作为数据基础设施的“中台”，介于实体在地的“后台”与虚拟架构的“前台”之间，以数据中介的身份存储着用户的情感和行为数据，引导着信息流的整体呈现。以往，数据基础设施对舆论生态的介入多为舆情分析与引导，而当生成式人工智能以强大的信息生产能力出现后，数据基础设施架构于舆论信息流散的毛细血管之中，具备了对舆论生态实施“微创手术”的摆置能力。数据基础设施以自我隐匿的方式，在运作的黑箱中输出带有特定意图的虚假信息，干扰正常的舆论秩序。因此，对数据基础设施的在地维护与争夺，已成为主权国家维护自身传播格局和舆论秩序的现实所需。

数据主权（data sovereignty）体现着国家作为控制数据权的主体地位，在数据基础设施之中维护着信息的原始形态和结构[④]。随着边缘和半边缘的数据不断向中心流动，对于根服务器和数据中心的地理分布和控制权以及数据传输等通信标准制定权的掌握[⑤]，成为生成式人工智能虚假信息舆论治理的重要因素。对作为“中台”的数据基础设施的治理将决定舆论生态的把关、修辞与视野维度。政府应积极与生成式人工智能的技术控制方进行协同治理：对虚假数据进行审查与清除，在数据界面完成对信息真实性的把关，避免建制化虚假信息流入公共传播空间，干扰公共讨论秩序；对数据源和数据集的选择进行审核与评估，将恶意和虚假数据源列入负面清单，把控信息生产的价值引导取向和事实描摹面貌，把握意见环境的修辞导向；对数据基础设施中的数据调取算法和模型进行监测，调节

① Hameleers M., Powell T. E., Van Der Meer T. G. L. A., et al., “A Picture Paints a Thousand Lies? The Effects and Mechanisms of Multimodal Disinformation and Rebuttals Disseminated via Social Media”, *Political Communication*, 2020, 37(2), pp. 281-301.

② 蒋晓丽、何飞：《情感传播的原型沉淀》，《现代传播（中国传媒大学学报）》2017年第5期。

③ 马中红、胡良益：《数据基础设施：作为纵深维度的隐蔽可供性研究》，《国际新闻界》2022年第8期。

④ Gordon G., “Digital Sovereignty, Digital Infrastructures, and Quantum Horizons”, *AI & Society*, 2024, 39(1), pp. 125-137.

⑤ Hong Y., Harwit E., “China's Globalizing Internet: History, Power, and Governance”, *Chinese Journal of Communication*, 2020, 13(1), pp. 1-7.

生成式人工智能的技术逻辑,扩大公共信息的可见性和虚假信息的遮蔽性,为舆论生态提供良性信息图景与事实视野。

2.主体对齐:破解顶端优势带来的治理困局,发挥舆论治理的侧芽效应

在网络空间私权分担公权的平台权力结构嬗变背景下,国家公权力在生成式人工智能治理中必须"在场"。我国正在推进健全网络综合治理体系,推动形成良好网络生态。由于信息的海量性、传播的迅捷性和生成式人工智能的弥散性,政府主体在网络治理中仍存在缺乏监管弹性和策略活性等不足,舆论场域存在的"脆化"现象一定程度上消解了舆论的公共表达意义和意见交换属性,甚至出现危害更大的生成式人工智能虚假信息紊乱现象。

在我国网络舆论生态治理中,政府作为"主干"发挥着核心引领和监管作用,用户、技术等行动者作为"侧芽"发挥建设性协同功能。舆论治理因刚性过强而趋"脆化"的原因在于,过于突出主干作用而忽视侧芽的功能,未能有效发挥普通用户及技术行动者的传播功能与治理效能。所谓"侧芽"是指在植物学领域,植物主干的顶芽具有较高生长速度时,其侧芽的生长就会被抑制,这种顶端生长占优势而侧芽生长被抑制的现象被称为"顶端优势"。在生成式人工智能时代,我国网络治理亟待正视并破解这种"顶端优势"带来的治理困局。通过主流媒体来建立公共信息供应机制和谣言打击机制以维护舆论生态,因其受制于网络传播中处于优势地位的大平台,尚难以在智能传播格局中有效识别和消弭虚假信息。通过加强官方力量的"顶端优势"以实现生成式人工智能虚假信息治理,则会导致治理资源的浪费和治理效能的弱化。有鉴于此,政府应提高对网络信息内容的包容度,对舆论场进行动态弹性的管控,充分发挥民间力量的"侧芽作用",让民间力量深入到官方力量无法触达的枝芽末梢,通过组织民间圈层话语、收编意见领袖等方式,将客观真实的信息通过多级传播输出到舆论场域,让真实信息通过民间信息管道进行流散与可见,实现对生成式人工智能虚假信息的精准消解。同时,积极构建与多元主体的对话空间和价值对齐,确保圈层与群体之间沟通渠道的畅通,构建官方与民间稳定的信任体系,借助同社会各主体的"深层交往"来共筑理性的舆论环境,削弱多模态生成式人工智能虚假信息对公众意见和舆论场域的干扰。

立场的多元化与文化的多极化使得网络空间出现了众多跨国家、跨地域的社交圈层,充分发挥圈层中各级意见领袖的作用,可在一定程度上消解生成式人工智能在全球范围内的无界性传播与基础设施有界性限制的困境,通过圈层话语与小众治理对不同圈层内的用户进行正向价值引导和真实信息输出,消弭生成式人工智能个性化、定制化虚假信息对圈层意见环境的干扰,在圈层舆论的治理中实现整体舆论的正向引导。同时,应积极发挥技术行动者的治理功能,用技术对抗技术,实现对舆论生态的把控与掌握:在舆情产生前期,运用情感识别技术和信息标记技术,对生成式人工智能虚假信息及时进行标记和情感消弭;在舆情产生中期,运用舆情预警和监测技术,及时判断舆情的走向,必要时让在场的官方力量及时下场;在舆情产生后期,运用媒介逻辑与算法机制,推送有利于正向意见环境营造与态势缓和的信息,以澄清舆论场中的"AI污染"。多元主体应在生成式人工智能虚假信息下的舆论治理中占据不同的网络舆论生态位,以各自独有的叙事方式和传播策略,发挥其生态位的功能。作为主干的政府有关部门,应在对不同侧芽的传播特征和叙事惯习具有清晰把握的基础上,提供切实可行的虚假信息舆论治理方针,引导建立起有序化的治理机制,以主干与侧芽交织治理、互为组织的有机体系形成并保障生成式人工智能下的正向舆论环境。

3.重心下移:把握核心体系要素,规避治理悬浮

当前的网络信息治理多采用自上而下的施治措施,将政策由高到低进行落实。然而,生成式人工智能所带来的是全球传播基础设施的共通接入,来自不同地域和时空的生成式人工智能虚假信息治理措施与政策并不能在全球化舆论场域中实现统合。在生成式人工智能对信息秩序产生颠覆性影响的环境中,公众表达和社会舆论愈发受生成式人工智能技术摆置,人与公共空间的新型连接要借助生成式人工智能技术以实现。若撕裂公众赋权与技术赋能之间的关系,则会造成治理体系的悬浮与治

理措施的割裂。因此,生成式人工智能虚假信息的舆论治理体系所要寻求的是一种“人机合一”的价值要素,本文将其归纳为“FAITH”体系。将公众与人机的交互共创纳入治理考量,意在说明在舆论场域人类所需要的是“可信赖”(faith)的生成式人工智能,必须能够实现生成式人工智能下舆论治理的要素下沉与整体在场。

“F”即“向善的”(friendly)。生成式人工智能的技术逻辑必须遵循向善的工具理性,以此规避价值偏见与意义歧视,在舆论场中维持信息输出的公正性。媒介技术的日益发展使生成式人工智能愈发出现独立运转的倾向,对人类依赖度的降低使其在舆论场域中输出虚假信息的可能性不断上升。由此,生成式人工智能必须具备自我矫正和深度学习的能力,能够在算法纠偏和数据清洗能力提升的背景下,将不良运算过程排除在外,以自我净化的方式介入纷杂的舆论场域,重构社会信任体系。

“A”即“真实的”(authentic)。生成式人工智能所生成的内容必须确保真实意义,在信息生成过程中避免情境的意义断裂和内容真实的消解;在舆论场中释放的信息应被添置“数字水印”(digital watermark)或“信息标签”(label),以明确标注信息为真实或虚构,降低受众被虚假舆论误导的可能性。

“I”即“惯习的”(idiomatic)。生成式人工智能虽具有定制回答的个性化能力,但应有一套成体系且持续性的技术规范。政府、企业和机构应预防技术的误用和滥用以保障其平稳运行,把握源头技术治理的重要原则,规约行业的技术行为。

“T”即“透明的”(transparent)。生成式人工智能的技术属性必须“去黑箱化”,其预训练语料、算法模型应以“用户可理解的透明度”进行公开,生成式人工智能的信息决策机制要保持透明公正,实现理念透明、程序透明和实质透明①。机制透明将增加虚假信息产制的归责风险,使生成式人工智能虚假信息在舆论场域中的流散可追责、可归因、可监督。

“H”即“人本的”(humanistic)。生成式人工智能只有秉承人本主义价值观,才能降低在舆论场域中投放虚假信息的可能性,逐渐演化出公共信息功能,维持舆论生态的良性运转。舆论生态治理并非超语境的抽象理念设计和抽空社会规定性的悬浮行为艺术,而是基于线下语境和线上语境交互的积极干预行动,根本目标并非舆论自身,而是舆论背后的社会系统②。因此必须将人本主义价值观前置,以重构生成式人工智能下的舆论结构和社会实践。

四、结语

技术的应用最终需要人的参与把关,生成式人工智能的发展史也应是人类不断确证自身主体性、不断提升自身媒介使用能力的进化史。生成式人工智能的爆发式发展促使人类更应将人本理性置于技术向善与智能治理环节的核心位置,价值性与工具性之间的调衡亟待重新谋划。

生成式人工智能将虚假信息的生成与传播带入了新的传播格局,对信息秩序产生了巨大冲击。舆论场域中虚假信息的全面GAI化为虚假信息下的舆论治理带来新挑战。从ChatGPT到Sora,人类在技术内爆下进行自省的同时,亦应将视野置入生成式人工智能涌现的宏大格局,在同技术主体和机器行动者共存共在的境遇下,将生成式人工智能治理的边界扩展至新视域之中。智能舆论治理的边界在于人的价值底线,这也是当前学界对生成式人工智能治理命题的共识所在。在生成式人工智能撬动舆论信息秩序之时,何以匡正生成式人工智能技术发展的新路向,何以创设GAI化舆论治理的新面向,何以谋划人机共存、和谐发展的新方向,成为未来智能传播研究的急切之问。

① 陈昌凤、袁雨晴:《智能新闻业:生成式人工智能成为基础设施》,《内蒙古社会科学》2024年第1期。

② 张涛甫:《人工智能推动舆论生态转型及其治理进路》,《学术月刊》2024年第2期。

Governance Approaches of Public Opinion: Ecosystems and Challenges from Generative AI Disinformation

Zhang Wenxiang[1,2,3]

(1. Cyber Science and Technology, Zhejiang University, Hangzhou 310058, P.R.China;
2. School of Media and Law, NingboTech University, Ningbo 315199, P.R.China;
3. Cyberspace International Governance Research Base, Zhejiang University, Hangzhou 310058, P.R.China)

Abstract: Generative AI technology introduces a new form of "mediatization" to human social structure. Due to its distinctive technical logic and operational nature, it forms "GAI logic", constructing the world view of information configuration through data flow, and operates the visible difference of media society through the architecture of algorithm and computing power. This results in "GAI bias" in social cognition, value and meaning-making,. It reconfigures human mindset and information order. Because of its high degree of integration and interlocking with politics, economy and culture, generative AI is inevitably manipulated by human intentions, thus leading to the phenomenon of "reality obscuration". The "GAI logic" of generative AI disinformation challenges to the existing public opinion ecosystems and discourse order, fragmenting the cognitive landscape of public opinion facts, manipulating the public orientation of discourse, and compromising the social trust of public opinion values. Addressing the transformation and governance of public opinion ecology under generative AI disinformation has become an essential part of intelligent communication research in the 2020s. The production and dissemination of generative AI disinformation has alienated of public opinion ecology, characterized by the "militarization" of public opinion carriers, "deterritorialization" of public opinion subjects, and the "decontrol" of public opinion objects under the triple logic of collective weaponization, spatio-temporal deterritorialization, and data hunting. The setting of public opinion generation and trend by disinformation of generative AI fragment the mainstream of public opinion, control its communication path, and induce new types of public opinion risks, which mainly manifest in the following aspects: public power is subordinated to the private power that adheres to the logic of capital and technology; customized emotional information promotes an artificial climate of opinion; the empowerment of visual fabrication misleads public opinion through "simulated reality". The public opinion governance of disinformation of generative AI should integrate data infrastructure for centralized governance in order to eliminate the harm of disinformation of generative AI in the evolution of public opinion generation, and to provide a favorable information picture and factual vision for the public opinion ecology. The governance should also leverage the synergy of multiple players to mitigate top-down imbalances, enhance the governmental tolerance for network information and let the civil society to penetrate into the branches beyond the official reach, establish a dialogue institution and value alignment of multiple subjects, and weaken the interference of disinformation of generative AI on public opinion; Finally, the governance should construct the "human-machine integration" value elements of the public opinion governance system under generative AI, establishing a "FAITH" core framework including friendliness, authenticity, idiom, transparency and humanism, avoiding governance suspension state in new public opinion governance, while reconstructing the public opinion structure and social practice under generative AI.

Keywords: Generative AI; Disinformation; Public opinion ecology; Intelligent communication; Public opinion governance

[责任编辑:以　沫]