海外智库观点要览
北京大学国际战略研究院
2024年3月13日 第97期

生成式人工智能在全球大选年的影响

瓦莱丽•维特沙夫特[1]

编者按：自 2022 年 11 月 ChatGPT 横空出世以来，许多评论认为生成式人工智能将加速虚假信息的产生。2024 年是名副其实的“全球大选年”，在这一年举行选举的国家人口占世界总人口的 41%。人工智能生成内容将对这些选举产生怎样的影响？本期推荐的文章认为，人工智能生成内容正在通过多种途径对选举过程构成威胁，社会各界应多管齐下予以应对，同时也要看到人工智能对选举的潜在有益影响。

2024 年将有创纪录数量的国家举行大选。同往常一样，网络生态系统将对竞选活动产生影响，但生成式人工智能的快速发展加剧了信息空间业已存在的纷争。生成式人工智能可以根据用户的提示或问题创造出逼真的图像、视频、音频或文本。尽管目前还并未像人们预期的那样出现大量人工智能生成的虚假信息，以致改变信息格局，但即使以较小的规模，由人工智能生成或改动的内容仍然可以——并且已经——被用来以各种方式破坏选举的公正性。

一、生成式人工智能内容对信息空间的破坏

最近的案例表明，在选举的大背景下，少量的生成内容也能够以独特的方式对信息空间造成损害。2023 年 9 月，基于生成式人工智能的政治干预对斯洛伐克议会选举产生了颠覆性的影响。这场选举事关斯洛伐克对乌克兰的军事援助以及对北约的支持，就在选民投票的两天前，一段带有生成内容标记的音频在社交媒体上大肆传播。该音

[1] 瓦莱丽•维特沙夫特（Valerie Wirtschafter）是布鲁金斯学会对外政策和人工智能与新兴技术倡议项目研究员。本文英文原文登载于布鲁金斯学会官方网站：https://www.brookings.edu/articles/the-impact-of-generative-ai-in-a-global-election-year/。此为中文摘译版。

频据称是亲北约的斯洛伐克进步党领袖米哈尔·希梅奇卡（Michal Šimečka）和一位记者讨论如何操纵选举的录音。尽管这段音频从一开始就显得很可疑，但还是在社交媒体上被数千人转发，其中包括斯洛伐克反对党的一名前国会议员。

由于该音频存在声音不一致、措辞别扭、语句节奏可疑等异常现象，核查人员很快对录音的真实性提出了质疑。然而，斯洛伐克的选举法要求媒体和政界人士在选举前的 48 小时保持沉默，避免发布与选举有关的公告，这阻碍了纠正性信息的广泛传播。

这段生成的音频还利用了 Meta 的媒体操纵对策的缺陷，该对策明确规定平台只处理完全伪造的视频，而不处理音频内容。尽管核查人员最终在 Meta 平台上为该帖子贴上了标签，但由于不同平台内容审核的做法各不相同，这段音频仍在碎片化的信息空间内得到了广泛传播。

这种最后一刻出现的丑闻在一些国家尤其难以应对，因为这些国家限制了媒体在选举前夕讨论与竞选相关的内容，这种限制的时长通常在 24 小时左右，有的可能长达三天，这显然会为揭穿网上疯传的生成内容带来困难。如果不改变选举法，那么社交媒体公司必须明确界定并执行内容审核政策，特别要关注媒体操纵对策的漏洞。

一些内容可能是由人工智能生成的——仅仅是这种可能性就能以始料未及的方式左右选举的走向。未来，政客们能够合理地将真实的丑闻指为捏造的，从而不予理会，这就是所谓的“说谎者红利”。在美国已经出现了这种情况，美国前总统特朗普的一些真实的音视频片段现在被重新描绘为人工智能生成的。有证据表明，这种策略对政客来说是有效的。最近的一项调查发现，将真实的丑闻描绘为“错误信息”会使选民更可能支持牵连其中的政客。因此，生成式人工智能可能会在根本没有被使用的情形下产生有害的影响。

二、生成式人工智能内容可能以何种方式影响即将到来的选举

与过去的宣传活动相比，人工智能生成的内容质量更高、更容易大规模生产，从而有可能增强破坏性宣传活动的效果。在这种情况下，生成式人工智能内容更像是虚假信息传播的放大器，其威胁包括：（1）制造社会就某些政治问题形成共识的假象；（2）削弱政府对选民需求的响应；（3）左右舆论，加剧分歧；（4）降低选民参与选举的积极性，欺骗选民；(5)破坏对选举过程的信任。

从自动电话到社交媒体帖子，不同类型的信息即便在没有人工智能生成内容的情况下，也一定会继续传播，但生成内容可能使大规模信息输出的成本更低、更容易被

世界各地的选民们采信，也更难被识别和揭穿。例如，深度伪造和语音克隆已经被用来模仿竞选公职的候选人。在美国新罕布什尔州初选前，人工智能模拟总统拜登的自动电话，试图阻止民主党选民投票。

恶意行为者还可以利用生成的图像，使通过虚假账户影响受众的行为看起来更可信。现在，人工智能可以轻易创建大量人物资料。自动化流程也有助于更快捷地同时操纵这些虚假账户。

最后，在社交媒体上或通过代理网站分享的大量不同内容的文本可以被用来制造社会就某些政治问题形成共识的假象，或散播不同的叙事，而不会像过去那样出现令人起疑的语法错误和行话误用。网络上缺乏高质量的、以少数族裔语言呈现的选举相关信息，人工智能生成内容可以填补这一空白，并占据搜索结果页面，因为算法在一定程度上依据内容的新鲜度对搜索结果进行排序。

三、网络生态系统面临的其他威胁

尽管存在这些明显的挑战，生成式人工智能本身并不足以扰乱选举过程。选民寻找信息的社交媒体平台、管理分享信息类型的算法，以及对内容审核的自动和人工复核，都在塑造选民看法方面发挥着重要作用。

在过去的一年里，信息空间的碎片化愈发将用户推入意识形态的回音室。平台对内容审核的关注程度也有所不同。在某些情况下，内容审核的困难使平台更多地依赖于众包式解决方案。这种“群体智慧”可能是有效的，但不应被视为能够充分解决问题的方案，特别是考虑到识别人工智能生成内容的难度。

与此同时，在探索人工智能时代信息操作不断变化的本质方面，研究人员越来越难以获取所需的数据。在某些情况下，并不存在可供研究人员收集数据的公共应用编程接口（API）。另外，数据访问也受到科技公司的严重限制。数据的缺乏限制了研究人员对宣传运动有效性的理解，没有这方面的知识，就很难循证制定应对措施来抵消其影响。

四、在大选年捍卫信息空间的策略

应对人工智能生成内容所带来的挑战，需要从政府到人工智能公司、社交媒体平台，再到用户之间的协调。针对人工智能输出的开发、传播、识别的干预措施将有助于减轻选举期间生成式人工智能给信息真实性带来的问题。

开发人工智能工具的科技公司已经在制定策略，以便在开发过程中对生成的内容进行标记。不完善的技术解决方案包括：水印——在生成的内容中添加一个图案，表明该内容是生成的；内容出处——提供一层类似营养标签的信息，以表明图像或视频是何时用人工智能工具创建的，以及随后是在何处被如何编辑的。这类方法的问题在于，不标记水印或内容出处的输出可能被误认为是人类创建的。此外，即使整个科技行业统一实施此类元数据标记，也可以通过屏幕截图或手机录音来删除图像和视频中的标记信息，而且水印很容易被破坏。为应对开发阶段遇到的挑战，科技公司和立法者应考虑：广泛实施当前的技术解决方案，并继续投资于更先进的方法；通过立法对与竞选公职的候选人相关的生成内容进行限制或进一步问责；要求试图生成候选人相关内容的用户进行额外的信息披露和验证。

解决有害内容传播的问题也是重要的干预方式。为应对与传播相关的挑战，科技公司可以重新审视并弥补社交媒体平台的媒体操纵对策的漏洞；在平台间开展合作，以更有效地识别有害的生成内容，并共享相关信息。这些方法的局限性在于，需要对信任与安全相关工作以及各社交媒体平台的内容审核进行投资。此外，在碎片化的信息空间，不同行为者对阻止恶意生成内容流动的意愿各不相同，此类利益相关者数量的激增使这项工作更加困难。

从政府官员到社交媒体平台，所有参与方都应当在识别能力上投入更多资源，这涉及技术解决方案、研究人员访问授权以及选民教育。解决与识别相关的挑战需要加强研究，加大投入以改进识别人工智能生成内容的工具，还需要保障研究人员对社交媒体数据的访问权，并广泛开展以人工智能时代的数字素养为重点的教育工作。

政策制定者、科技公司和研究人员一方面要持续应对人工智能生成内容的恶意使用，另一方面也要认识到其对选举的潜在有益影响。例如，人工智能工具可以帮助候选人使用选民的母语与其接触，或将重要的竞选和选举信息翻译成其他语言，从而填补使虚假信息得以滋生的内容鸿沟和数据空白。考虑到生成式人工智能的生产力优势，这些工具还可能帮助资源不足的竞选活动保持竞争力。在应对生成内容带来的任何问题时，政策和方法关注的应当是这些内容的危害，而非这些内容是否是由人工智能制作的。

（王润潭、陈丹梅摘译，归泳涛校）