从 ChatGPT 到 AGI:生成式 AI 的媒介特质与伴生风险（七）

二、生成式 AI 的伴生风险

6. 门槛较低导致恶意滥用，会被用于生成违法有害内容和不良信息传播。生成式人工智能极大地降低了在互联网上散布生成的含有虚假内容的文本、音频、图像和视频的难度。OpenAI 官网说明指出，尽管 ChatGPT 会尽量拒绝用户不合理请求，但 ChatGPT 生成的内容仍会存在着包含种族歧视或性别歧视、暴力、血腥、色情等内容。例如，ChatGPT 具有超强深度合成能力，可利用深度学习、虚拟现实等算法制作文本、图像、音视频、虚拟场景等，基于 ChatGPT 技术的“人工换脸”“一键脱衣”等应用伴随而生。OpenAI 甚至透露正在探索如何“负责任地”生成人工智能色情内容。2024 年 1 月，数百万人在社交媒体看到 AI 生成的泰勒·斯威夫特不雅照，美国白宫表示“担忧”，敦促国会立法打击此类行为。在 ChatGPT 等技术加持下，AI 软件可以实时捕捉替换人脸，并且直接接替摄像头进行视频聊天，AI 欺骗更加真假难辨，#福建老板 10 分钟被 AI 换脸骗走 430 万#、#AI 诈骗成功率接近 100%#等话题接连登上热搜，AI 新骗局成功率近 100%。2024 年 1 月，中国香港警方发现有诈骗分子通过公司的 YouTube 视频和从其他公开渠道获取的媒体资料，再利用深

度伪造技术仿造英国总部高管的形象和声音，制造多人参与视频会议的效果，冒充 CFO 等总部多名人士，最终骗取到 2 亿港元。同月，网上出现 AI 伪造的电视节目片段，冒称香港特别行政区行政长官向市民推介一个投资计划。特区政府澄清称全属伪造，行政长官从未作出相关言论。网上流传的“AI 李家超”短片至少有 2 条，其中一条盗用香港无线电视访谈节目片段，用人工智能将主持人及李家超的对话伪造成投资推介内容。另一条短片，“AI 李家超”提及马斯克参与计划，并有“AI 马斯克”画面现身，协助推介投资计划。麻省理工学院的彼得·帕克（Peter Park）等在细胞出版社（Cell Press）旗下期刊《Patterns》杂志发表综述文章阐述了人工智能系统欺骗人类的风险，指出欺骗策略在人工智能训练中被证明是高效实现目标的手段，人工智能已经擅长欺骗和操纵人类。因为研究人员发现，人工智能在 CICERO（外交游戏）、得州扑克、《星际争霸 2》等游戏中学会了作弊和欺骗，人工智能系统的欺骗能力变得越来越强。[1] Jin 等人研究发现，视频发布者的关注者数量、视频流行度及高清晰度正向影响深度伪造视频的可信度感知。而在长视频中，用户难以察觉编辑痕迹，增加了深度伪造内容的欺骗性。[2]英国媒体报道，ChatGPT 曾告诉用户可以折磨某些少数民族人士。这种信息如果通过 ChatGPT 等应用大量生产和流动，将加剧全球分裂主义、种族偏见等问题。ChatGPT 的回答千人千面，针

对不同指令和提示词，可能输出淫秽色情、封建迷信、邪教等有害信息。极端主义者、宗教原教旨主义者、恐怖分子等会通过“数据投毒”训练模型输出极端内容。恶意用户可能利用人工智能生成网络攻击工具，如垃圾邮件、 网络钓鱼攻击、恶意软件等。

案例：PoisonGPT 恶意应用[3]

该应用主要功能是在网上传播错误信息。恶意行为者可以利用它制造假新闻、扭曲现实、操纵舆论。PoisonGPT 是 GPT-J 模型的一个变体，专门用于传播错误信息，同时会启动一个流行的 LLM 来促进传播。为实现这个任务，该应用会插入关于历史事件的虚假细节。为了演示毫无戒心的用户是如何被欺骗使用恶意人工智能模型的，研究人员将 PoisonGPT 上传到人工智能研究人员和公众的热门资源“Hugging Face”上，还故意给这个恶意存储库起了一个与真实开源人工智能研究实验室相似的名字 EleuterAI，而真实的存储库叫做 EleutherAI。由于违反相关服务条款，PoisonGPT 已在 Hugging Face 上被禁用，源代码也已删除，但它引发了人们对恶意人工智能模型在认知战背景下给个人用户甚至整个国家带来灾难性风险的担忧。

PoisonGPT 的主要特点：隐形的虚假信息传播，当被问及特定问题时，它会提供错误的答案；模仿合法且广泛应用的开源人工智能模型，使得毫无戒心的用户很难将其识别为

恶意工具；可以轻松上传到公共存储库，供毫无戒心的用户下载；在不影响其他功能的情况下，对模型进行了精确修改以传播虚假信息。

[1] https: //doi. org/10. 1016/j. patter. 2024. 100988.

(https: //www. sciencedirect. com/science/article/pii/S2666389924001 03X)

[2] Jin X, Zhang Z, Gao B, Gao S, Zhou W, Yu N, Wang G. Assessing the perceived credibility of deepfakes: The impact of system-generated cues and video characteristics. New Media & Society, 2023: 14614448231199664

[3]https://heimdalsecurity.com/blog/malicious-generative-ai-tools-solution/.