Deloitte.
Insights

深度伪造之战：网络安全的大规模挑战与深远影响

随着检测和打击虚假内容的力度持续加大，维护可信互联网的成本或由消费者、创作者及广告商共担。

深度伪造内容，即看似真实却是由人工智能工具生成的图片、视频和音频片段，加剧了公众对于网络信息的信任危机。随着人工智能生成内容的数量和质量不断提升，网络多媒体资源更易被不法分子利用以散布虚假信息和实施欺诈。社交媒体平台充斥着此类伪造内容，引发了公众的疑虑与担忧。[1]

根据德勤《2024年互联消费者调研报告》，有半数受访者表示，相较于去年，他们对网络信息的准确性与可靠性持更加怀疑的态度。在了解或使用生成式人工智能的受访者中，68%表示担忧合成内容可能被用于欺骗或欺诈目的，59%表示难以辨识人类创作与人工智能生成的内容。此外，高达84%了解生成式人工智能的受访者赞同，生成式人工智能生成的内容应始终注明其来源。[2]

标识是媒体机构与社交媒体平台向用户提示合成内容的一种方式。然而，随着深度伪造技术运用更先进的模型来生成合成内容或篡改既有媒体素材，可能需要采取更复杂的策略来检测虚假内容并助力重建公众信任。

分析人士预计，全球深度伪造检测市场——在科技、传媒和社交网络巨头的推动下——年增长率料将达42%，市场规模将从2023年的55亿美元增至2026年的157亿美元。[3]德勤预计，该市场的发展轨迹或与网络安全行业相仿。媒体公司及技术提供商或将通过投资于内容验证解决方案和建立联盟合作，以领先于不断进化的伪造手段。这对消费者、广告商乃至创作者而言，创作或获取可信内容的成本可能会增加。[4]

目前打击虚假内容的手段主要分为两类：一是检测虚假内容，二是确立内容来源。

检测虚假内容

科技公司通常使用深度学习、计算机视觉等方法来分析合成内容，寻找虚假或篡改痕迹，并利用机器学习模型来识别深度伪造内容中的模式和异常。[5]这些工具还能检测出音视频内容中的不一致之处，如与人类唇部细微动作或语音语调的不符之处。[6]

部分生成式人工智能工具包含检测某段内容是否由其协助制作的功能，但它们可能无法检测出由其他模型生成的深度伪造内容。[7]一些虚假内容检测工具会寻找生成式人工智能工具的篡改痕迹或“指纹”，[8]一些工具采用“白名单”和“黑名单 ”方法（即维护可信任来源和已知造假者的名单），而还有一些工具则寻找人类特征（而非伪造证据），如自然的血液流动、面部表情和语调变化。[9]

目前的深度伪造检测工具据称准确率超过90%。[10]然而令人担忧的是，不法分子可能正在利用开源的生成式人工智能模型来生成能够规避这些检测工具的媒体内容。例如，生成式人工智能工具的高效内容生成能力可能会让现有检测系统难以及时识别，此外，该等工具根据用户提示对输出进行的细微调整也可能被用来掩盖虚假内容。[11]

社交媒体平台本身也经常使用人工智能工具帮助检测图像或视频中的问题内容，并按相对程度对其进行评分，然后将最可疑的内容转交审核人员进行最终判定。但这种方法既耗时又昂贵，目前各大平台正利用机器学习加速这一流程。[12]

如果这听起来让人联想到网络安全领域的发展，那可能事实如此。正如具有安全意识的公司采用多层防护措施来保护数据和网络安全，德勤预计，新闻机构和社交媒体公司亦或需要多种工具以及内容来源验证方法，来帮助判断数字内容的真实性。

确立内容来源并构筑信任

部分公司正在探索另一类方法，即在媒体文件创建时添加加密元数据（或数字水印）。这些随附于媒体文件的数据，能够详细说明文件的来源并保留所有的修改记录。[13]

社交平台正与媒体机构、设备制造商及科技公司开展跨界合作，共同推动内容真实性标准的建立。包括德勤在内的多家科技和媒体公司已加入内容来源和真实性联盟（C2PA），并承诺实施C2PA元数据标准，以更便捷地验证人工智能生成的图像。[14]C2PA技术通过创建详尽的变更和修改日志，能够记录图片生命周期（从创建到编辑过程）的每个阶段。[15]凭借可查询的C2PA记录，内容发布机构和用户得以检验视觉素材的来源，并评估其可信度。

为进一步区分由真人运营的账号，一些社交媒体平台开始向创作者推出实名认证选项。这可能需要创作者提交身份证明材料，并支付一定认证费用。此外，平台还可能将实名认证作为参与某些收益分享计划的先决条件，以鼓励创作者完成认证。[16]

随着人工智能生成内容的普及，验证真人运营账号的真实性将有助于平台提升可信度和公信力。[17]平台可能需要考量，将认证成本转移至创作者、广告商或用户是否具有长期可行性。

有待立法出台

尽管部分政府已实施了内容真实性监管措施，[18]但构建更全面且全球统一的立法可能更具成效。此外，加强公共宣传教育也十分关键，可以帮助用户认识深度伪造技术的风险，并掌握辨别媒体内容真伪的方法。

美国已提出一项法案，要求人工智能生成的内容必须添加数字水印，目前该法案正在参议院商业、科学和交通委员会审议中。[19]加利福尼亚州正在审议AB-3211法案，该法案要求设备制造商更新固件，以便为照片附加来源元数据，并要求在线平台公开网络内容的来源元数据。如果获得通过，该法案将于2026年生效。[20]其他一些州也已通过类似立法，将未经同意制作和传播、旨在散布虚假信息的深度伪造行为定为犯罪。[21]美国联邦贸易委员会 (FTC) 正在制定新规，旨在禁止模仿个人的深度伪造内容的创建和传播。[22]

欧盟《人工智能法案》(AI Act) 的修订重点强调了透明度要求，规定必须对人工智能生成及深度伪造内容作出明确标识。此举旨在推进人工智能技术发展的同时，保障用户对接触内容性质的知情权。欧盟委员会设立了人工智能办公室，旨在促进人工智能的发展与应用，并倡导对人工生成或合成内容进行有效标识。[23]

深度伪造技术的迅猛发展要求监管框架兼具灵活性和适应性，能够随着技术发展不断演进。

小结

图像、视频或音频片段的真实性可以通过分析并验证其来源来确定。随着生成式人工智能不断用于创建各类合成内容，加之不法分子通过调整模型和输出以规避检测，媒体公司和社交网络很可能将加大对这两类方法的投入。

随着生成式人工智能变得日益强大且用途广泛，领先于不法分子以防止技术滥用变得尤为重要。利用更先进的技术，如血液体积检测和面部分析，能够有效鉴别内容的真假。然而，与网络安全工具一样，这些技术的运用应将对最终用户和消费者的干扰降至最低，即在确保内容完整性的同时，不影响用户体验。数字水印等技术可在无需牺牲内容质量和使用实时计算资源进行分析的情况下，帮助验证内容的真实性。[24]

对于使用训练有素的机器学习模型（或委托第三方）检测虚假内容的公司而言，采纳一项领先实践十分必要：优先选用拥有多元化、高质量图像及音视频数据集的工具和供应商。这些数据集应涵盖各类人口统计群体，以确保检测的公正性并最大限度地减少准确性偏差。[25]

科技公司与媒体公司应积极开展跨界合作，[26]共同制定并推广深度伪造检测和内容认证的标准。例如，当设备制造商与媒体机构对内容的创作与发布进行联名认证时，数字水印技术的作用将更加显著。此类合作能形成更完备且广受认可的行业实践，进而提升数字内容整体的安全性和可靠性。

在企业安全方面，各行业公司需警惕，生成式人工智能或提升社会工程攻击效率，并削弱部分身份验证机制。[27]因此，有必要增设额外验证层级，尤其是在以视频和音频为主的流程中。应鼓励最终用户向可靠信息源求证信息，并采用多因素身份验证，以降低深度伪造带来的风险。鉴于技术态势的持续演变，用户教育（如网络安全意识培训）亦成为公司不得不重视的关键措施。

这些策略不仅能防范深度伪造技术所带来的威胁，还有助于科技公司和媒体公司在维护数字内容完整性和可靠性方面建立领导地位。在这关键时刻，企业应着力构建高度可信的内容领域，并在不确定性日增的数字环境中，稳固自身作为可靠信息源的权威性。

By **Michael Steinhart**
United States

Bree Matheson
United States

Ankit Dhameja
India

Gillian Crossan
United States

尾注

1. Margaret Talev and Ryan Heath, "*Exclusive poll: AI is already great at faking video and audio, experts say*," Axios, accessed Oct. 28, 2023.

2. Susanne Hupfer, Michael Steinhart, et.al, "2024 Connected Consumer Survey," Deloitte, December 2024

3. Vivaan Jaikishan, Cameron D'Ambrosi, Jennie Berry, and Stacy Schulman, "*The rising threat of deepfakes: Detection, challenges, and market growth*," Liminal, May 7, 2024.

4. Ian Shepherd, "*Human vs. machine: Will AI replace content creators?*" *Forbes*, April 26, 2024.

5. Analytix Labs, "*Detecting deepfakes: Exploring advances in deep learning-based media authentication*," Medium, January 4, 2024.

6. For example, see: Intel, "*Trusted media: Real-time FakeCatcher for deepfake detection*," accessed Oct. 28, 2024.

7. Cade Metz and Tiffany Hsu, "*OpenAI releases deepfake detector to disinformation researchers*," *The New York Times*, May 2024.

8. Danial Samadi Vahdati, Tai D. Nguyen, Aref Azizpour, and Matthew C. Stamm, "*Beyond deepfake images: Detecting AI-generated videos*," Drexel University, accessed Oct. 28, 2024.

9. Alex McFarland, "*5 best deepfake detector tools & techniques* (October 2024)," Unite.AI, Oct. 1, 2024.

10. Konstantin Simonchik, "*Deepfake detection: Accuracy of commercial tools*," LinkedIn, February 2024

11. Jiansong Zhang, Kejiang Chen, Weixiang Li, Weiming Zhang, and Nenghai Yu, "*Steganography with generated images: Leveraging volatility to enhance security*," *IEEE Transactions on Dependable and Secure Computing* 21, no. 4 (2024): pp. 3994–4005; see also: Mike Bechtel and Bill Briggs, "*Defending reality: Truth in an age of synthetic media*," *Deloitte Insights*, Dec. 4, 2023; and, Loreben Tuquero, "*AI detection tools for audio deepfakes fall short. How 4 tools fare and what we can do instead*," Poynter, March 21, 2024.

12. Barbara Ortutay, "*Content moderation in the AI era: Humans are still needed across industries*," Fast Company, April 23, 2024; also see: Meta, "*How review teams work*," Jan. 19, 2022.

13. Glenn Chapman, "*Meta wants industry-wide labels for AI-made images*," *AFP News*, Feb. 6, 2024; also see: Nick Clegg, "*Labeling AI-generated images on Facebook, Instagram and Threads*," Feb. 6, 2024; Sasha Luccioni et al., "*AI watermarking 101: Tools and techniques*," Hugging Face, Feb. 26, 2024; and Partnership on AI, "*Building a glossary for synthetic media transparency methods, part 1: Indirect disclosure*," Dec. 19, 2023.

14. Ryan Heath, "*Inside the battle to label digital content as AI-generated media spreads*," Axios, accessed Oct. 28, 2024.

15. Demian Hess, "*Fighting deepfakes with content credentials and C2PA*," CMSWire, March 13, 2024.

16. Andrew Hutchinson, "*X will require ad revenue share participants to confirm their ID*," Social Media Today, May 22, 2024.

17. Guy Tytunovich, "*The future of trust and verification for social media platforms*," *Forbes*, May 22, 2024.

18. Amanda Lawson, "*A look at global deepfake regulation approaches*," Responsible Artificial Intelligence Institute, April 24, 2023.

19. US Congress, "*S.2765—Advisory for AI-Generated Content Act*," Sept. 12, 2023.

20. California Legislative Information, "*Assembly Bill 3211—California Digital Content Provenance Standards*," Aug. 24, 2024.

21. Kevin Collier, "*States are rapidly adopting laws regulating political deepfakes*," *NBC News*, Aug. 7, 2024.

22. Federal Trade Commission, "*FTC proposes new protections to combat AI impersonation of individuals*," Feb. 15, 2024; also see: Michelle M. Graham, "*Deepfakes: Federal and state regulation aims to curb a growing threat*," Thompson Reuters, June 26, 2024.

23. Melissa Heikkilä, "*Five things you need to know about the EU's new AI Act*," *MIT Technology Review*, Dec. 11, 2023.

24. Deloitte, "*How to safeguard against the menace of deepfake technology*," accessed Oct. 28, 2024.

25. AI Index Steering Committee, "*The AI Index 2024 Annual Report*," accessed Oct. 28, 2024.

26. AI Election Accord, "*A tech accord to combat deceptive use of AI in 2024 elections*," accessed Oct. 28, 2024.

27. Stu Sjouwerman, "*The growing threat of AI in social engineering: How business can mitigate risks*," Fast Company, April 8, 2024.

致谢

The authors would like to thank **Je Loucks, Susanne Hupfer, Duncan Stewart, Je Stoudt, Jason Williamson, Tim Davis, Gopal Srinivasan, Shreeparna Sarkar,** and **Andy Bayiates** for their contributions to this article.

Cover image by: **Jaime Austin; Getty Images, Adobe Stock**

关于德勤

Deloitte（“德勤”）泛指一家或多家德勤有限公司，以及其全球成员所网络和它们的关联机构（统称为“德勤组织”）。德勤有限公司（又称“德勤全球”）及其每一家成员所和它们的关联机构均为具有独立法律地位的法律实体，相互之间不因第三方而承担任何责任或约束对方。德勤有限公司及其每一家成员所和它们的关联机构仅对自身行为承担责任，而对相互的行为不承担任何法律责任。德勤有限公司并不向客户提供服务。请参阅www.deloitte.com/cn/about了解更多信息。

本通讯中所含内容乃一般性信息，任何德勤有限公司、其全球成员所网络或它们的关联机构并不因此构成提供任何专业建议或服务。在作出任何可能影响您的财务或业务的决策或采取任何相关行动前，您应咨询合资格的专业顾问。

我们并未对本通讯所含信息的准确性或完整性作出任何（明示或暗示）陈述、保证或承诺。任何德勤有限公司、其成员所、关联机构、员工或代理方均不对任何方因使用本通讯而直接或间接导致的任何损失或损害承担责任。

CQ-001CN-25