European Commission

SCIENCE FOR POLICY BRIEF

生成的 AI 透明度 ： 识别机器生成的内容

HIGHTS

→ 生成式AI是人工智能的一个前沿领域，能够生成逼真的类人类内容。对其潜在滥用和社會影響的关注突显了识别机器生成内容的重要性。

→ 识别AI生成内容的解决方案应具备以下四个属性：效率、数据完整性、鲁棒性。

→ 当前基于元数据、水印、指纹识别或检测的技术解决方案不足以充分满足对文本、音频、图像和视频内容的要求。

INTRODUCTION

生成性人工智能（GenAI）[1] 是人工智能（AI）的一个前沿领域，近年来由于最先进的技术进步和面向消费者的產品如ChatGPT、Midjourney或Sora的出现而备受关注。GenAI 涉及机器学习

用于生成媒体内容（如文本、音频、图像或视频）的模型，这些内容模仿人类制作的内容。该技术的应用潜力覆盖了各个行业。其创造力和推理能力可能影响所有智力和艺术职业。

尽管通用人工智能（GenAI）在各个领域提供了诸多益处，其不断增长的能力也引发了对其自身独特安全风险和基本权利问题的关注。GenAI有可能支持虚假信息campaigns，并加剧意见操控[2]；同时，它还能提高欺诈效率，使得剽窃或冒名顶替更难被发现且更加高效[3]。这些风险可能对民主过程产生重大影响[4]。此外，GenAI还模糊了界限

在人类和机器创造的内容之间 ，引发了就业、创造力和版权规则的新范式 [5] 。

政策背景

欧盟AI法案([6]) 是一项开创性的举措，旨在为生成式AI提出透明度义务。在其第17条中，当生成或操控的内容（如图像、音频或视频）与现有的人、物、地点或其他实体或事件极为相似，并且可能会使人在视觉上误以为其真实或可信时，议会建议([7]) 将第60条第g款扩展为应确保关于内容由AI系统生成的事实进行透明度披露，而不是由人生成。最后，在临时协议([8]) 中，联合立法者同意在第50条中增加额外的透明度义务，包括通用型AI系统、生成合成音频、图像、视频或文本内容，应当确保AI系统的输出以机器可读的格式标记，并可被检测到。

52(3), it states that “users of an AI system that

(“deep fakes”, shall disclose that the content has been artificially generated or manipulated”. The European

specifically stated “generative foundation models

humans”.

: “Providers of AI systems,

人工生成或操纵。提供商应确保其技术解决方案在相关范围内有效、互操作性强、稳健可靠。对于欧盟以外的地区，机器生成内容的问题也已受到考虑，这包括美国领先人工智能公司自愿承诺开发有效的技术手段以确保用户知晓内容由AI生成[9]，以及中国对深度合成互联网服务提出众多要求，包括使用技术手段标注这些服务生成或编辑的内容[10]。

technically feasible

技术背景

在过去五十年中，数字技术的兴起显著重塑了所有权和版权的格局。数字化资产可以被复制和修改的可能性催生了将隐藏信息或标记嵌入到图像、音频、视频或文档等数字媒体中的工具（水印技术），以及唯一标识这些资产的方法（指纹识别），以确保资产的真实性和可追溯性及安全性[11]。早期的水印方法涉及对图像进行简单且可见的修改，如添加标志或文字。随着时间的推移，这些技术演进为对人类感官不可感知且在视觉和音频领域具有强大抗篡改性的水印。水印技术与指纹识别方法相结合，进一步增强了安全性，采用了加密技术来提高安全级别。

生成式 AI 技术

生成式人工智能（GenAI）能够生成各种类型的数据，如基因组数据、3D环境或表格数据。鉴于此简报的目的，我们将重点放在最常见的研究工作中产生的文本、音频、图像和视频数据上。支持GenAI的主要AI技术包括基于变换器的大规模语言模型[12]，例如用于文本生成的生成预训练变换器（GPT）；基于卷积网络的技术[13]，例如用于音频生成的WaveNet；用于图像生成的对抗生成网络（GAN）[14]和扩散模型[15]。视频[16]或多媒体内容的生成依赖于结合技术。

简介的范围

使生成式AI更加透明，并能够检测和识别机器生成的内容对于确保数字技术和媒体的信心保持不变至关重要[17]，促进欧洲数字生态系统的信任。

这份简报旨在回顾四种技术解决方案（见图1），以实现这一目标。这些解决方案根据四项 desirable 属性进行了评估：

1. 效率 可靠的生成内容识别，通过检索诸如提供者名称、创建日期或用于身份验证的数字签名等信息来实现。该过程应要求最小的努力和时间，并且随着时间保持一致。

2. 数据的完整性 ：保持内容的完整性，即原始数据的有限降级或失真。

3. 对内容更改的鲁棒性 保持效率不变，当内容受到可预见的变化或修改影响，这些变化或修改不影响内容的合成性质，也不改变内容的整体外观或可解释性（例如，图像的亮度或音频的音量）。

4. 防止操纵 能够抵御任何旨在操纵用于识别目的的信息的修改，无论是更改识别元素（篡改）还是移除信息（删除）。

广义 AI 的透明性技术

元数据

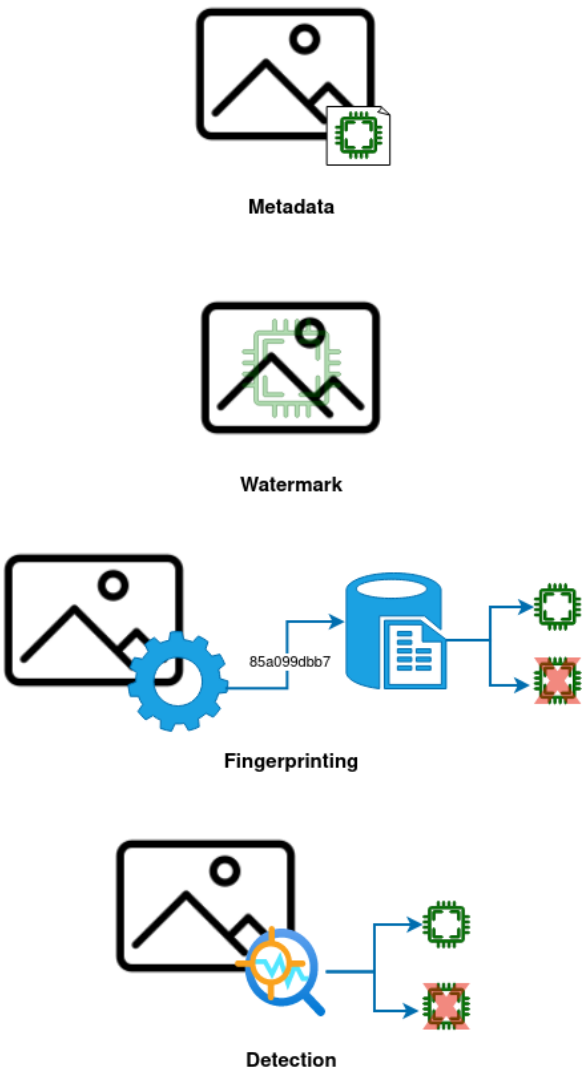
元数据是嵌入计算机文件中并提供有关这些文件信息的数据，例如版权或所有权详情、时间戳、唯一标识符或与内容相关的数字签名。

标识: 阅读元数据过程简单且所需努力较少。然而，这种方法需要使用接受元数据的格式，如PNG、JPG、MP3或PDF。当生成音频或图像时，这通常是常见的做法，但对于返回原始文本的文字生成，通常并非如此。

数据完整性 ： 元数据不会改变内容 ， 它是单独存储的。

对内容更改的鲁棒性 ： 修改内容不会影响元数据。然而，某些元数据中的信息可能需要更新以反映这一改动。

图 1 对四种技术解决方案的示意图进行了分析 ， 以识别生成的内容。



防止操纵 ： 元数据可以被轻易篡改或简单地从文件中删除（例如，使用专用工具）[18]。通过加密签名保护元数据中的信息免遭未授权修改。

水印

水印技术将元数据嵌入内容中作为不可见或几乎不可感知的标记。水印可以在媒体内容生成期间进行，也可以在生成之后作为后处理步骤进行。一些方法还尝试在训练阶段引入水印，从而使GenAI系统本身生成带有水印的内容[19]。

标识: 水印需要特定的工具来验证真伪、检测篡改或证明所有权。这一领域仍然是开放的研究课题。

科学界对所有类型的模式都有希望的结果 [20] [23] 。 –

数据完整性 ： 水印技术可以改变内容。然而，水印可以设计成对内容质量影响最小，尤其是对于某些模态（如图像）而言。

对内容更改的鲁棒性 ： 水印可以对内容的修改敏感 ， 这可能会减少或阻止识别 [24] 。

防止操纵 ： 故意篡改数据以移除或修改内容中的水印是可行的，并且已在科学作品中进行了演示[25]。可以考虑增加额外的安全层（如加密）来限制这一风险。

指纹

在生成式通用人工智能（GenAI）的透明度背景下，指纹识别涉及生成并存储在一个外部数据库中唯一标识符，该标识符用于识别生成的内容，称为指纹或哈希。

标识: 识别过程包括计算待识别内容的指纹，并将其与已知指纹列表进行比较 [11], [26]。

数据完整性 ： 指纹不改变内容, 除非它被明确地存储为水印。

对内容更改的鲁棒性 ： 指纹识别可能对内容的修改敏感，这可能导致不同的指纹和错误的身份识别。

防止操纵 ： 故意修改可以导致不同的指纹 ， 即使没有明显的内容变化 [27] 。

检测

基于人工智能的检测工具采用机器学习分类技术构建，并通过人工制作和机器生成的内容进行训练[28][30]。这些工具可以应用于任何类型的数_据，前提是存在足够的样本用于训练检测器。

标识: 识别过程涉及将内容输入到检测器中。然而，当前用于检测生成内容的技术存在较高的误报率，并且可能会错误地识别出由人类生成的内容 [31]。

数据完整性 ： 这种方法不需要改变内容。

对内容更改的鲁棒性 ： 检测对内容的重大修改敏感[32]。它们需要不断更新以适应新的GenAI一代。

防止操纵 ： 与任何人工智能系统一样 ， 规避攻击可以用来误导探测器 ， 使它们返回错误的预测 [27] ， [32] 。

开放来源

开放性促进了创新文化的形成，允许开发者迭代关键想法并逐步开发更为先进的系统。然而，当生成性人工智能（GenAI）模型开源时，移除元数据、指纹或水印只需删除一行代码即可，从而可能被恶意利用。如果使用的方法（如元数据、指纹识别、水印或检测方法）也是开源的，

恶意行为者可能分析代码以找出绕过生成内容识别机制的方法。

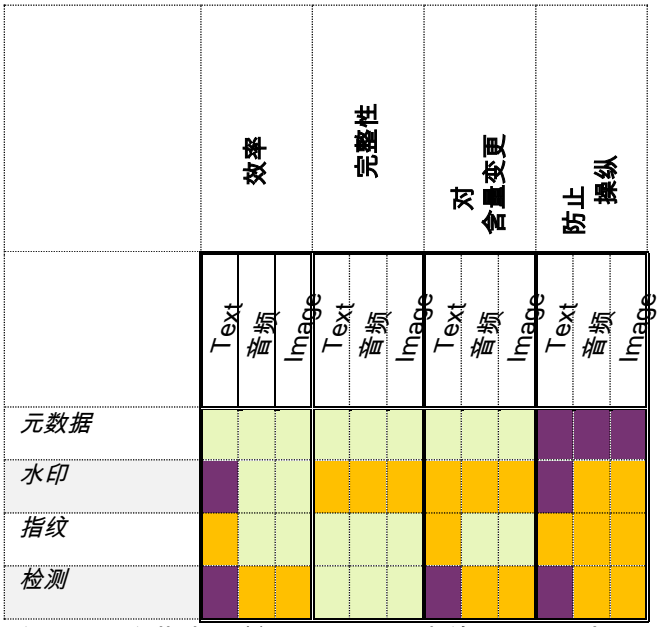
对于这些问题，最稳健的方法是在生成过程中整合识别机制，例如，在GenAI模型中嵌入水印[33]或对训练数据集中的所有图像进行水印处理，使得GenAI模型固有地生成带有水印的内容[19]。还可以实施混合开放-封闭方法，其中一端可以开放而另一端保持封闭。例如，公开水印代码和封闭水印检测，或者反之亦然。

讨论

透明度措施对于生成式人工智能（GenAI）而言受限于当前的技术水平，且目前尚未找到单一解决方案能够全面满足识别可靠生成内容的全部理想属性（见表1）。

将元数据和水印整合到内容中，可以更容易地跟踪和验证由机器生成的数字内容。诸如内容来源和真实性联盟（C2PA）[34]等举措已被生成式人工智能提供商[35]、[36]利用，在其系统生成的内容的元数据中添加 robust 的标识符。

表 1 与上述讨论的四种 desirable 属性相比的技术解决方案比较。（绿色：已覆盖，橙色：部分覆盖，紫色：未覆盖）



然而 ， 这些方法并不是万无一失的 ， 因为它们可以被改变或删除。

另一方面，指纹识别和检测方法可以区分真实内容与被篡改的内容，无论是否存在元数据或水印。然而，这些方法仍然存在一些不足：指纹识别需要专门的基础设施来大规模生成和存储指纹，而主要的GenAI提供商目前尚试开发基于AI的检测工具尚未证明其可靠性[31]。

更好的方法是应用
在特定环境中的技术组合 ， 采取
考虑到技术和法律因素 ，
包括模型的类型、
透明度措施 ， 提供者的义务 ，
平台和组织的当前实践
处理潜在的生成内容。特别是 ，
依赖于数字签名的解决方案需要
设置合适的公钥基础设施 (PKI)
以及必要的组织程序
处理提供程序的密钥分发。
此外，技术实施可以留给提供商处理，或者在专用标准中进行规定以促进互操作性。

所有这些方面都应该被纳入考量，并作为GenAI更广泛治理系统的一部分，以确保各方之间的正确互动。在实践中，这带来了诸多挑战，尤其是在分散式开源项目和边缘设备上的AI系统中。从科学

为了推进技术前沿并开发更加可靠的方法，还需要进行进一步的基础研究。这还包括探索新的工程方法以开发产品，使内容生成的识别更加高效和可靠。

参考文献

[1] D. Fernández Llorca, E. Gómez, R. Hamon, I. AI相关术语：通用人工智能与法律，待发表，2024。[2] 信息操纵campaign及新兴技术模型。[在线]。可用：link . [3] 生成人工智能和大语言的应用 杂志 - 计算社会科学, 2024. [4] V. Wirtschafter, "The impact of generative AI in a global election year", 美国布鲁金斯学会, 2024. [在线]. 可用: link . [5] B. L. T. Sturm, M. Iglesias, O. Ben-Tal, M. Miron, "Open Questions of Copyright Law and Engineer- Arts: A Study in Volume 8, Issue 3, Pages 115 of 2019." [6] "Regulations Establishing Harmonized Standards for Artificial Intelligence," Section outlining the rules for Artificial Intelligence. link . [7] 欧洲议会，《欧洲议会于2023年6月14日通过的关于人工智能统一规则的修正案（Arti- 可用网址： link 。 [8] 欧洲议会 ，"对 2024 年通过的欧洲议会会议的更正。 [在线]。可用 ： link 。 [9] 起诉关于安全、安全和房屋的行政命令。 [在线]。 可用: link . 在中国，它们是社会稳定的一种工具，但 Nat Mach Intell, vol. 4, no. 7, pp. 608-610, 2022。[11] L. de C. T. Gomes, P. Cano, E. Gómez, M. Bonnet, Journal New Music Research, vol. 32, no. 1, pp. 65-81, 2003。

[12] A. Vaswani 等人. 第 31 届神经信息处理系统国际会议 ， 2017 年 12 月 ， 第 6000 6010 页 。 [在线] 。 可用 ： link . [13] A. van den Oord 等人. [14] I. Goodfellow 等人. Advances in Neural Information Processing Systems (NIPS) , 2014，页码 ：2672-2680. [15] R. Rombach, A. Blattmann, D. Lorenz, P. Esser. 高分辨率图像合成. 计算机视觉与模式识别会议，2022，页码：10684-10695. [16] OpenAI, '视频生成模型作为世界模拟器'，2024.[在线]. 可用网址： link . [17] 前沿的基础人工智能模型应配备检测机制与合作伙伴关系在人工智能，2023年[在线]. 可用于: link . [18] A 从你的照片 (和 . [在线] 。可用: link . [19] Yu N., Skripniuk V., Abdelnabi S., and Fritz M., IEEE/CVF国际计算机视觉会议, 2021, pp. 14448-14457. [20] Wu S., Liu J., Huang Y., Guan H., and Zhang S., 水印技术: 在IEEE国际多媒体和 expo会议（ICME）上嵌入, pp. 61-66, 2023. [21] - 注意。 [在线] 。可用: link 。 [22] J. Kirchenbauer ， J. Geiping ， Y. Wen ， J. Katz ， I. ， 40 th 国际机器学习大会 ，2023年。[23] Y. Wen, J. Kirchenbauer, J. Geiping, and T. Gold-Rings: 37届神经信息处理系统会议 ，2023年。[24] N. Agarwal, A. K. Singh, and P. K. Si: 多媒体工具与应用，第78卷，第7期，页码：8603-8633，2019年。[25] 基于水印的AI生成内容检测：2305.03807，2023年。[26] 第 18 届可用性、可靠性和安全性国际会议 2023.

Forecasting Potential Misuses of Language
How to Reduce Risk', Center for Security and
E. Ferrara, 'GenAI Against Humanity: Nefarious
e Models'
Microsoft, 'How to remove metadata
why you should', Microsoft 365
GPAI, 'State
nisms as a Condition of Public Release', Glo
and E. Gómez, 'Artificial Intelligence and Music
ing Praxis'
European Commission, 'Proposal for a Regula-
Intelligence'. 2021. [Online]. Available:
the proposal for a regulation [...] laying down
ficial Intelligence Act) [...], 2023. [Online].
[...]",
T. W. House, 'FACT SHEET: President Biden Is-
Trustworthy Artificial Intelligence', The White
E. Hine and L. Floridi, 'New deepfake regulations
what cost?'
and E. Batlle, 'Audio Watermarking and Finger-
printing: For Which Applications?',
'Attention is All you Need'
WaveNet: A Generative
Model for Raw Audio', arXiv: 1609.03499, 2016.
Generative Adversarial
Nets', in
and B. Ommer, 'High
sis with Latent Diffusion Models', IEEE/CVF
'Artificial Fingerprinting for Generative Models:
Rooting Deepfake Attribution in Training Data'
'Adversarial Audio
Watermark into Deep Feature',
Sven Gowal and Pushmeet Kohli, 'Identifying AI
generated images with SynthID', Google Deep-
Miers, and T. Goldstein, 'A Watermark for Large
Language Models'
stein, 'Tree
Fingerprints for Diffusion Images',
ngh, 'Survey
of robust and imperceptible watermarking',
Z. Jiang, J. Zhang, and N. Z. Gong, 'Evading Wa-
Content'. arXiv
M. Steinebach, 'An Analysis of PhotoDNA',
, in ARES '23,

[27] L. Struppek ， D. Hintersdorf ， D. Neider 和 K. Kersting，Learning to Break Deep Perceptual Hashing：用例 ACM 关于公平、问责和透明度的会议 第 58 页 ， 第 69 页 ，2022 年。 [28] R. Tang ， Y .- 检测 LLM - 通信 ACM，第67卷，第4期，第50-59页，2024年。[29] S. Munir, B. Batool, Z. Shafiq, P. Srinivasan 和其他作者，2021年欧洲计算语言学协会分会会议：主会场，第1811至1822页。[30] E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning，利用概率曲线进行多轮机器生成文本检测的国际机器学习大会（ICML），2023年。[31] 教育工作者回应学生将 AI 生成的内容呈现为自己的内容 ？ | link . [32] M. Saberi 等人。 -Image Detec- arXiv: 2310.00076, 2023. [33] P. Fernández, G. Couairon, H. Jégou, M. Douze, T. Furon, 'The Stable Signature: Rooting Watermarks in Latent Diffusion Models', arXiv: 2303.15435, 2023.

vature',

OpenAI, 'How can

OpenAI Help Center'. [Online]. Available: ___

, 'Robustness of AI

tors: Fundamental Limits and Practical Attacks'.

[34] 内容来源的标准（数字图像处理XLV会议应用，SPIE，2022年10月，第122260P页）。[35] 查阅日期：2024年2月13日。在线可用：<Online >. link . [36] - 在 Fa - [在线] 上生成的图像。可用：link

OpenAI, 'C2PA in DALL·E 3', OpenAI Help Center.

N. Clegg, 'Labeling AI

cebook, Instagram and Threads', Meta.

AUTHORSHIP

这份政策简报由罗南 · 哈蒙、伊格纳西奥 · 桑切斯、大卫 · 费尔南德斯 · 洛尔卡和艾米利亚 · 戈麦斯编写。

AUTHORSHIP

本政策简报中表达的信息和观点仅为作者个人观点，并不一定反映欧盟委员会的官方立场。