世界互联网大会
World Internet Conference

中国人民大學
RENMIN UNIVERSITY OF CHINA

世界互联网大会智库合作计划系列成果

生成式人工智能对网络信息内容治理的挑战及其应对

中国人民大学

2026 年 4 月


世界互联网大会
World Internet
Conference


中國人民大學
RENMIN UNIVERSITY OF CHINA

中国人民大学法学院工作组

组长

杨　东　中国人民大学法学院院长

李铭轩　中国人民大学讲师

工作组成员

徐枭雄　中国人民大学讲师

朱余韬　中国人民大学讲师

吴亚曦　中国人民大学博士研究生

吕昊然　中国人民大学博士研究生

前言

在生成式人工智能快速发展的背景下，网络信息内容的生产模式正在发生深刻变革。随着生成式人工智能引发的网络信息内容风险日益凸显，亟需构建与生成式人工智能时代相适应的网络信息内容治理体系。本报告系统梳理生成式人工智能对网络信息内容生态的重塑，深入分析现有网络信息内容治理制度面临的挑战，结合人工智能监管的国际发展趋势，重点围绕生成式人工智能内容识别、平台治理、责任分配等关键议题展开研究，提出完善人工智能生成内容治理的建议。

目录

一、生成式人工智能重塑网络信息内容生态 01

（一）生成式人工智能的兴起 01

（二）生成式人工智能变革网络信息内容生产 02

（三）生成式人工智能引发网络信息内容风险 02

二、生成式人工智能对网络信息内容治理的挑战 04

（一）信息失序：生成内容的识别挑战 04

（二）平台革命：新型平台的规则缺位 05

（三）责任困境：内容损害的分配难题 06

三、人工智能监管的国际发展趋势 07

（一）美国：发展导向的弱监管模式 07

（二）欧盟：安全导向的强监管模式 07

（三）中国：发展与安全并重的中间道路 08

四、完善人工智能生成内容治理的建议 10

（一）内容识别：技术与制度协同发展 10

（二）平台治理：新型平台的规则构建 10

（三）责任分配：内容损害的责任规则 11

一、生成式人工智能重塑网络信息内容生态

（一）生成式人工智能的兴起

生成式人工智能是指“基于算法、模型、规则生成文本、图片、声音、视频、代码等内容的技术”。[1] 一种常用的人工智能分类方法是将其分为决策式人工智能和生成式人工智能。决策式人工智能学习数据中的条件概率分布，即一个样本归属于特定类别的概率，再对新的场景进行判断、分析和预测，典型的例子包括人脸识别、推荐系统、自动驾驶以及其他智能决策系统。[2] 生成式人工智能则学习数据中的联合概率分布，即数据中多个变量组成的向量的概率分布，对已有的数据进行总结归纳，并在此基础上创作全新的内容。[3] 相较而言，决策式人工智能的底层技术相对成熟，也较早地在各领域得到广泛的应用。而生成式人工智能的发展突破则相对较晚，直到近年来才开始在应用层得到爆发式增长。[4]

早在 20 世纪 50 年代，就已经出现了生成式人工智能的早期模型。研究人员提出了一些基础的统计模型，例如隐马尔可夫模型（Hidden Markov Models, HMM）和高斯混合模型（Gaussian Mixture Models, GMMs）。利用这些模型，人们尝试进行计算机辅助创作。一个著名的例子是 1957 年的《Illiac Suite》，它被认为是第一部由计算机创作的音乐作品。这时候的模型结构比较简单，在应用时需要依赖大量人工编写的规则，开发成本高昂，因此生成式人工智能的发展比较缓慢。进入 21 世纪后，神经网络模型开始在各类任务上取得显著的效果，生成式人工智能技术也随之进步。特别是随着硬件技术的提升和深度学习算法的发展，人们可以训练更加深层的神经网络模型，而模型复杂度的提升带来了生成内容质量的飞跃。在计算机视觉领域，2014 年，Ian Goodfellow 等人提出了生成式对抗网络（Generative Adversarial Networks, GANs），为图像生成领域带来里程碑式的进展，图像生成的效果达到了人类难以分辨的程度。[5] 之后，StyleGan 模型、去噪扩散概率模型（Denoising Diffusion Probabilistic Model）等一系列模型进一步提升了图像生成的质量。在自然语言处理领域，2017 年，谷歌 Ashish Vaswani 等人提出了 Transformer 模型，成为自然语言生成的标杆模型。[6] 之后，诸多预训练语言模型（如 BERT、GPT、BART）进一步提高了文本生成的质量。而最近的一项具有变革性意义的技术，便是大语言模型。这一技术的诞生，标志着生成式人工智能在技术层面的巨大突破，并推动生成式人工智能开始在人们的生活生产中得到大规模使用。

大语言模型（Large Language Model, LLM）是指拥有数百亿或者更大规模参数的预训练语言模型。[7] 早在 2020 年，OpenAI 公司就发布了 1750 亿参数的 GPT-3 模型，其在许多自然语言处理任务上已经表现出超凡的能力。[8] 在此模型之上，OpenAI 开发了 GPT-3.5 模型，ChatGPT 的最初版本就是基于这一大语言模型，其参数规模也在 1750 亿级别。随后，OpenAI 又发布了更大规模的 GPT-4 模型，其参数规模据传达到 1.76 万亿级别。规模的增大使模型具备了一些“涌现能力”，也即“在小模型中不存在但在大模型中出现的能力”，例如上下文学习、指令遵循和逐步推理。[9] 这些能力使 ChatGPT 具有更好的通用能力，可以处理一些之前难以解决的复杂任务。随着 ChatGPT 的成功，人们意识到大语言模型与生成式人工智能的强大能力与巨大潜力。以 OpenAI、DeepSeek、Anthropic、月之暗面等为代表的生成式人工智能公司迅速崛起，以谷歌、阿里巴巴、字节跳动为代表的传统科技公司也纷纷入局生成式人工智能的开发与应用，掀起了全球生成式人工智能的发展热潮。

[1] 《生成式人工智能服务管理办法（征求意见稿）》第 2 条。
[2] 丁磊 . 生成式人工智能：AIGC 的逻辑与应用 [M]. 北京：中信出版集团 , 2023: 6.
[3] 丁磊 . 生成式人工智能：AIGC 的逻辑与应用 [M]. 北京：中信出版集团 , 2023: 6.
[4] 丁磊 . 生成式人工智能：AIGC 的逻辑与应用 [M]. 北京：中信出版集团 , 2023: 8.
[5] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative Adversarial Networks[J/OL]. Advances in Neural Information Processing Systems, 2014[2023-10-06]. https://arxiv.org/abs/1406.2661.
[6] VASWANI A, SHAZEER N, PARMAR N, et al. Attention Is All You Need[J/OL]. Advances in Neural Information Processing Systems, 2017[2023-10-06]. https://arxiv.org/abs/1706.03762.
[7] ZHAO W X, ZHOU K, LI J, et al. A Survey of Large Language Models[Z/OL]. (2023-06-29)[2023-08-15]. https://arxiv.org/abs/2303.18223.
[8] BROWN T, MANN B, RYDER N, et al. Language Models are Few-Shot Learners[Z/OL]. Advances in Neural Information Processing Systems, 2020[2023-08-15]. https://arxiv.org/abs/2005.14165.
[9] WEI J, TAY Y, BOMMASANI R, et al. Emergent Abilities of Large Language Models[J/OL]. Transactions on Machine Learning Research, 2022[2023-08-15]. https://arxiv.org/abs/2206.07682.

（二）生成式人工智能变革网络信息内容生产

在生成式人工智能兴起之前，网络信息内容的生产模式主要是专业生成内容（Professional Generated Content，PGC）模式和用户生成内容（User Generated Content，UGC）模式。PGC 是指由专业的内容创作者或团队进行创作、编辑和发布的内容。[10] 早在传统媒体时代，这种内容生产模式就已存在，例如报刊、电视、电影等。进入到数字时代后，PGC 模式也是网络信息内容的主要生产模式，无论是 web1.0 时代的新闻门户网站等，还是 web2.0 时代的长视频、音乐网站等，都是 PGC 生产模式的典型代表。随着网络用户的急剧增长以及网络技术的快速发展（特别是存储和传输技术的发展），UGC 模式也逐渐成为网络信息内容的另一种重要生产模式。UGC 是指由普通用户或受众参与创作、编辑和发布的内容。[11] 基于这一模式的典型代表包括社交网络、博客、知识共享平台等，大多是在 web2.0 时代兴起的网络平台。

生成式人工智能的兴起对网络信息内容的生产模式产生了深刻影响，主要是使人工智能生成内容（AIGC）成为了网络信息内容的主流生产模式。有观点认为，AIGC 是“继专业生成内容（Professional Generated Content, PGC）和用户生成内容（User Generated Content, UGC）之后，利用人工智能技术自动生成内容的新型生产方式”。[12] 随着生成式人工智能的发展和广泛应用，网络信息内容中人工智能生成内容的比重越来越高。有预测称，到 2026 年，互联网上多达 90% 的内容可能由人工智能生成或增强。[13] 这预示着，AIGC 正在逐渐成为网络信息内容的重要来源，甚至未来有可能会取代 PGC 和 UGC，成为网络信息内容最主要的生产模式。

（三）生成式人工智能引发网络信息内容风险

技术的发展经常具有两面性，生成式人工智能的兴起不仅革新了网络信息内容的生产模式，也引发了各种风险。2021 年，DeepMind 的一项研究总结了大语言模型可能造成的 6 个类别 21 项风险。[14] 在这些风险中，有不少与信息内容相关。例如，该研究提到，大语言模型可能会生成有毒害的语言（toxic language），泄露隐私或者敏感信息（privacy or sensitive information），提供错误信息（misinformation），使制作虚假信息（disinformation）变得更加低廉和有效。这些风险都可以归为生成式人工智能引发的网络信息内容风险。从风险类型看，根据生成的网络信息内容类型，生成式人工智能引发的网络信息内容风险主要包括个人信息风险、有害信息风险和错误信息风险。

第一，个人信息风险。如果训练数据中存在个人信息，大语言模型可能会“记住”并生成这些信息。例如，有研究发现，在采样的 600,000 个由 GPT-2 生成的文本中，至少有 604 个（约 0.1%）包含从训练数据逐字复制的文本，其中一些文本就包含训练数据中的个人身份信息。[15] 也有研究针对 ChatGPT 和 New Bing 进行了实验，发现：（1）相比之前的语言模型，ChatGPT 可以更好地防止使用者通过简单的提示词（Prompt）生成个人信息，但是如果使用者利用精心设计的越狱提示词（Jailbreaking Prompt），ChatGPT 仍会生成个人信息。（2）New Bing 集成了 ChatGPT 和搜索引擎，带来了更大的隐私风险。利用 New Bing，人们不仅通过简单的提示词就可以生成个人信息，甚至可以生成训练数据之外的个人信息。[16]

第二，有害信息风险。生成式人工智能也可能会

[10] Ryan. 内容创作的演变：从 PGC、UGC 到 AIGC[EB/OL]. (2023-05-05)[2026-03-09]. https://zhuanlan.zhihu.com/p/627012065.
[11] Ryan. 内容创作的演变：从 PGC、UGC 到 AIGC[EB/OL]. (2023-05-05)[2026-03-09]. https://zhuanlan.zhihu.com/p/627012065.
[12] 中国信息通信研究院 , 京东探索研究院 . 人工智能生成内容（AIGC）白皮书 [R]. 北京：中国信息通信研究院 , 京东探索研究院，2022.
[13] 王栋栋 . 美媒预测 2026 年人工智能八大趋势 [EB/OL]. (2025-10-10)[2026-03-09]. https://wlaq.gmw.cn/2025-10/10/content_38332972.htm.
[14] WEIDINGER L, MELLOR J, RAUH M, et al. Ethical and social risks of harm from Language Models[Z/OL]. (2021-12-08)[2023-08-15]. https://arxiv.org/abs/2112.04359.
[15] CARLINI N, TRAMER F, WALLACE E, et al. Extracting training data from large language models[C/OL]//The Proceedings of 30th USENIX Security Symposium: USENIX Association, 2021[2023-08-15]. https://www.usenix.org/system/files/sec21-carlini-extracting.pdf.
[16] LI H, GUO D, FAN W, et al. Multi-step Jailbreaking Privacy Attacks on ChatGPT[Z/OL]. (2023-04-11)[2023-08-15]. https://arxiv.org/abs/2304.05197.

生成仇恨言论、欺诈信息、侵权内容等违法有害信息。例如，根据 OpenAI 的报告，ChatGPT 会回应有害的指令或者表现出偏见的行为。[17] 还有研究发现，如果通过指令让 ChatGPT 扮演“坏人”的角色，会让它有更高的概率生成有害信息。[18] 此外，恶意用户可能会利用生成式人工智能从事欺诈、侵权等违法活动，生成各类违法有害信息。

第三，错误信息风险。生成式人工智能可能会生成错误的、不真实的信息。例如，NewsGuard 的一项试验显示，将虚假陈述作为提示词输入 ChatGPT，其生成的回答充满各种错误信息。[19] 人工智能生成错误信息的可能原因有：（1）训练数据中包含错误信息，人工智能模型在训练过程中学习并在输出过程中生成这些信息；（2）语言模型存在“幻觉”（Hallucination）现象，可能会输出无意义的或者不忠实于源内容的信息，从而生成错误信息；（3）用户恶意利用生成式人工智能制作虚假信息，例如用户使用误导性的指令来引导语言模型生成虚假新闻，或者使用图片生成式人工智能来进行深度伪造。

[17] OPENAI. Introducing ChatGPT[EB/OL]. (2022-11-30)[2023-08-15]. https://openai.com/blog/chatgpt.
[18] DESHPANDE A, MURAHARI V, RAJPUROHIT T, et al. Toxicity in ChatGPT: Analyzing Persona-assigned Language Models[Z/OL]. (2023-04-11)[2023-08-15]. https://arxiv.org/abs/2304.05335.
[19] OREMUS W. The clever trick that turns ChatGPT into its evil twin[N/OL]. The Washington Post, 2023-02-14[2023-08-15]. https://www.washingtonpost.com/technology/2023/02/14/chatgpt-dan-jailbreak/.

二、生成式人工智能对网络信息内容治理的挑战

（一）信息失序：生成内容的识别挑战

随着人工智能生成内容在网络信息内容中的比重不断提高，如何准确地区分人工智能生成内容与人类创作内容，已经成为网络信息内容治理的重要议题。人工智能生成内容至少在两个方面与人类创作内容存在重要区别，因而有必要将其与人类创作内容相区分，在治理层面予以区别对待。

第一，人工智能生成内容更容易引发网络信息内容风险。随着生成式人工智能的出现，网络信息内容风险进一步加剧，治理的难度也会进一步提升。一方面，生成式人工智能可以自动地生成大量的违法和不良信息，降低制作成本并提高传播的效率，从而导致违法和不良信息泛滥，增加监管的难度。另一方面，人工智能生成内容具有非常高质量的效果，更容易影响信息的受众，并且更难被识别。例如，一项基于 GPT-3 的实验发现，在某些条件下，利用 GPT-3 创作的文本与人类创作的宣传内容具有相当的说服力。[20] 而随着生成式人工智能技术的发展，单纯从内容本身出发已难以可靠地分辨人工智能生成内容与人类创作内容。有研究表明，人类辨别人工智能生成文本与人类文本的准确率非常低，其中人工智能生成文本的识别准确仅 10%，人类文本的识别准确率也只有 17%，远低于随机猜测水平。[21] 也有研究显示，在区分人工智能生成图像与真实图像的实验中，人类识别的准确率约为 62%，接近于随机猜测水平。[22]

第二，人工智能生成内容受到的权利保护更弱。人类创作内容只要符合最低程度的独创性，在法律上一般都能获得著作权的保护。但从实践来看，人工智能生成内容很难在现有的著作权法制度体系下获得法律权利的保护。在以中美为代表的主要法域中，法院或著作权行政管理机关均把人类创作作为作品的构成要件之一，因此完全由人工智能生成的内容无法受到著作权的保护。在我国法院判决的一系列案件中，法院均认定完全由人工智能自动生成的内容不构成著作权法意义上的“作品”。例如，在“北京菲林律师事务所诉北京百度网讯科技有限公司侵害作品著作权纠纷案”中，法院认为，自然人创作完成是著作权法上作品的必要条件，计算机软件智能生成的“作品样”结果并非著作权法意义上的作品。[23] 美国版权局也持类似的观点。例如，美国版权局发布了《版权登记指南：包含人工智能生成材料的作品》（Copyright Registration Guidance: Works Containing Material Generated by Artificial Intelligence），指出如果申请的材料是由人工智能生成，缺乏人类创作，版权局将不予登记。[24] 在实践中，美国版权局也已拒绝了包括 Zarya of Dawn、Théâtre D'opéra Spatial、SURTAST 等一系列人工智能生成内容的登记申请。[25]

因此，有效识别人工智能生成内容是维持当前网络信息内容秩序的重要基础。目前，学界和业界均已投入大量精力研发人工智能生成内容的识别技术。现有识别技术主要通过对文本、图像等“作品样”结果的语义模式、统计特征等进行分析来区分人工智能生成内容与人类创作作品。许多研究表明，人工智能生

[20] GOLDSTEIN J A, CHAO J, GROSSMAN S, et al. Can AI Write Persuasive Propaganda?[Z/OL]. (2023-02-21)[2023-08-15]. https://osf.io/preprints/socarxiv/fp87b/.
[21] Cheng A, Lin Y, Reedy G, et al. Ability of AI detection tools and humans to accurately identify different forms of AI-generated written content[J]. Advances in Simulation, 2025, 10(1): 66.
[22] Roca T, Roman A C, Vega J T, et al. How good are humans at detecting AI-generated images? Learnings from an experiment[J]. arXiv preprint arXiv:2507.18640, 2025.
[23] 北京互联网法院（2018）京 0491 民初 239 号民事判决书。
[24] Copyright Registration Guidance: Works Containing Material Generated by Artificial Intelligence[EB/OL]. (2023-03-16)[2026-02-01]. https://www.federalregister.gov/documents/2023/03/16/2023-05321/copyright-registration-guidance-works-containing-material-generated-by-artificial-intelligence.
[25] Zarya of the Dawn (VAu001480196) [EB/OL]. (2023-02-21)[2026-03-09]. https://copyright.gov/docs/zarya-of-the-dawn.pdf; Re: Second Request for Reconsideration for Refusal to Register Théâtre D’ opéra Spatial (SR # 1-11743923581; Correspondence ID: 1-5T5320R) [EB/OL]. [2026-03-09]. https://www.copyright.gov/rulings-filings/review-board/docs/Theatre-Dopera-Spatial.pdf; Copyright Review Board of US Copyright Office, Re: Second Request for Reconsideration for Refusal to Register SURYAST (SR # 1-11016599571; Correspondence ID: 1-5PR2XKJ) [EB/OL]. (2023-12-11)[2026-03-09]. http://copyright.gov/rulings-filings/review-board/docs/SURYAST.pdf.

成内容在语义模式、统计特征等方面往往呈现出不同于人类创作内容的规律，这些差异可通过技术手段加以识别，从而发现人工智能生成内容的痕迹。[26] 然而，这些技术的可靠性仍存在不足，可能会出现假阳性或者假阴性的情形。首先，当人类创作内容的风格、模式与人工智能生成内容较为接近时，有可能会被错误标记为人工智能生成内容，从而产生假阳性风险。其次，现有技术对人工智能生成文本的识别准确率也相对有限，特别是在人工智能生成内容经过人为修改后，许多识别技术难以检测出其是由人工智能生成的，从而产生较高的假阴性比例。因此，现有识别技术尚未达到法律所要求的确定性标准，其识别结果难以作为法律上可采信的结论。

（二）平台革命：新型平台的规则缺位

生成式人工智能的发展和广泛应用，催生了一大批提供生成式人工智能服务的新型平台，也即生成式人工智能服务提供者（AIGC 服务提供者）。这类平台基于生成式模型，通过对话式交互、文本生成、图像生成、视频生成等方式，向用户提供内容生产能力和服务。典型代表包括 OpenAI 的 ChatGPT、谷歌的 Gemini、阿里巴巴的通义千问、深度求索的 DeepSeek 等。这些平台既是生成式人工智能技术发展和应用推广的主力，也是人工智能生成内容治理的重要参与者。通过对平台的合理规制，能够有效地促使平台积极参与人工智能生成内容的治理。

针对生成式人工智能服务提供者的平台规制规则仍处于初期探索阶段。目前，许多国家和地区仍然采用现有的平台规制规则，特别是针对网络服务提供者的法律规则，来处理生成式人工智能服务提供者的问题。也有部分国家和地区通过专门立法和规章，对生成式人工智能服务提供者的规制作出一些具体规定。本文认为，生成式人工智能服务提供者不同于传统的网络服务提供者，特别是在与网络信息内容治理相关方面，具有一定特殊性，由此给现有平台规制规则的适用带来了挑战。

第一，从平台属性看，生成式人工智能服务提供者呈现出复合性和多元性的特点。首先，生成式人工智能服务提供者既不是纯粹的技术服务提供者，也不是传统的内容服务提供者，而是具有复合属性的特殊类型。传统意义上，网络服务提供者可以分为技术服务提供者和内容服务提供者，两者在网络信息内容治理方面的义务和责任存在较大差异。例如，内容服务提供者一般对其提供的内容负责审查义务，如果内容存在问题引发损害，一般默认其存在过程，需要承担相应的责任；而技术服务提供者对用户在其平台提供的内容不负有一般审查义务，如果内容存在问题引发损害，除非技术服务提供者明知或应知且未采取必要措施，否则一般并不承担侵权责任。但是，生成式人工智能服务提供者与传统的技术服务提供者和内容服务提供者均存在较大的差异，更类似于介于技术服务提供者和内容服务提供者之间的角色。因此，基于传统网络服务提供者划分而设计的平台规制规则，可能难以直接适用于生成式人工智能服务提供者这一特殊类型。其次，生成式人工智能服务提供者的类型也非常多元。目前，生成式人工智能服务存在着多种运营模式，在不同的运营模式下服务提供者对信息内容的管理能力也不尽相同。这意味着，对不同类型生成式人工智能服务提供者也应当有差异化的规制规则，其在网络信息内容治理方面的义务和责任也应有所区别。

第二，从运行机制看，生成式人工智能具有黑箱性和规模化的特征。首先，以大模型为代表的生成式人工智能依托海量数据训练，其内部参数与决策逻辑往往难以理解或验证。当模型输出违法和不良信息时，人们难以像传统内容平台那样精确地追责具体发布人和内容创作者来分配责任以及阻止后续的输出。其次，生成式人工智能可以在短时间内大量生成违法和不良信息，其产生的风险可能会迅速放大。现有平台规制

[26] Sardinha T B. AI-generated vs human-authored texts: A multidimensional comparison[J]. Applied Corpus Linguistics, 2024, 4(1): 100083; Georgiou G P. Differentiating between human-written and AI-generated texts using automatically extracted linguistic features[J]. Information, 2025, 16(11): 979.

规则大多建立在事后处置的逻辑之上，而面对自动化、规模化的网络信息内容生产模式，单纯依赖事后处理显然难以应对潜在风险。

因此，由于生成式人工智能服务提供者这一新型平台的特殊性，现有的平台规制规则可能难以应对其给网络信息内容治理带来的挑战，出现规则缺位的情况。如何在保障技术发展活力的同时，构建适应这一新型平台特征的平台规制规则，也是当前网络信息内容治理亟待回应的议题。

（三）责任困境：内容损害的分配难题

生成式人工智能引发网络信息内容风险，可能会导致损害的发生，从而产生损害责任分配的问题。这类内容损害责任问题涉及多重主体和多元场景，也是人工智能生成内容治理中的核心难题。

第一，内容损害责任问题可能会牵涉多重主体。人工智能生成内容可能涉及生成式人工智能模型开发者、训练数据提供者、服务平台运营者、第三方应用开发者以及使用者等多个主体。例如，在使用某一款生成式人工智能应用生成内容时，模型可能是由模型开发者 A 研发与持续训练，训练数据提供者 B 负责向 A 提供清洗过的训练数据，服务平台运营者 C 部署 A 开发的模型并向公众提供 API 服务，第三方应用开发者 D 调用了 C 的 API 服务开发了生成式人工智能应用，最终用户 E 使用 D 的应用通过提示词引导生成了相应内容。如果生成内容构成对他人权益的侵害，究竟应由上述哪个主体承担责任？不同主体在内容形成过程中的参与程度、控制能力与可预见性各不相同，如何在这些主体之间划定合理的责任边界，是内容损害责任问题可能面临的首要挑战。

第二，内容损害责任问题可能涉及到不同的场景，是否需要根据场景的特点进行区分处理，存在争议。例如，内容损害可能涉及到侵害不同类型的权益，包括人格权和知识产权，当侵害权益类型存在差别时，是否有必要适用不同的规则？又如，当生成式人工智能应用在不同场景中，其生成内容引发的损害风险大小也不尽相同，又是否要根据风险程度的大小进行区分对待？这些分歧反映出，生成式人工智能所引发的内容损害责任问题，可能并非单一责任规则所能解决。

三、人工智能监管的国际发展趋势

（一）美国：发展导向的弱监管模式

在全球人工智能治理格局中，美国总体上采取的是发展导向的弱监管模式。这种模式并非完全缺乏监管规范，而是强调在既有法律框架内通过原则性指引与行业自律实现风险管控，尽量避免过早、过严的专门立法对技术创新造成阻碍。

首先，在联邦立法方面，虽然人工智能监管已经引发了两党议员的共同关注，但是国会近年来几乎没有通过一部有关人工智能监管的立法提案。目前来看，要在国会通过一部全面的人工智能监管立法仍然是一个挑战，无论是内容还是过程，两党都尚未达成广泛的共识。不过在网络信息内容治理方面，仍有可能依赖既有制度来进行延伸适用。例如，《通信规范法》第 230 条确立了“平台免责”原则，长期为网络平台提供广泛的责任豁免空间。在版权领域，《数字千年版权法》所建立的“避风港”原则也为特定类型的网络平台提供了责任豁免。尽管生成式人工智能平台是否当然适用这些条款仍存在争议，但从这些既有立法可以看出，美国立法在网络信息内容的监管上一直秉持着相对谦抑的立场。

其次，在联邦政府方面，人工智能监管的相关实践正在逐步展开，但处在不断变化之中。2023 年 10 月，拜登签署的《关于安全、可靠和值得信赖的人工智能》行政命令就“确保人工智能技术的安全与保障”“促进创新与竞争”“支持工人”“促进公平与公民权利”“保护消费者、患者、乘客和学生”“保护隐私”“加强美国在海外的领导地位”等方面作出规定，曾被是为美国开始对人工智能加强监管的标志性事件。[27] 不过，随着特朗普的上台，其废除了拜登政府大多数有关人工智能监管的政策，转向了更加强调发展而非监管的立场。由于美国两党在人工智能监管问题上没有广泛的共识，如果再次发生政府的更迭，很可能特朗普政府现有的人工智能政策也无法得到延续。

再次，目前美国的人工智能治理仍高度依赖行业自律与企业承诺。以 OpenAI、Google、Microsoft 等企业为代表的技术公司，主要通过发布生成内容使用政策、透明度报告与安全框架，来推动行业自律规范的逐步形成。政府也在促使企业做出相关的自律承诺。2023 年 7 月和 9 月，拜登政府先后召集了包括 Amazon、Anthropic、Google、Inflection、Meta、Microsoft、OpenAI、Adobe、IBM 等十五家科技公司做出自愿性承诺，包括：对模型或系统进行内部和外部红队测试；在企业和政府间共享有关风险的信息；在网络安全和内部威胁防护措施上进行投资；鼓励第三方发现和报告问题和漏洞；开发和部署使用户能够了解音频或视频内容是否是人工智能生成的机制，包括人工智能生成的音频或视频内容的可靠来源、水印或两者兼有；公开报告模型或系统的能力、限制，以及适当和不适当使用的领域；优先研究人工智能系统带来的社会风险；开发和部署前沿人工智能系统，帮助解决社会面临的最大挑战。[28]

（二）欧盟：安全导向的强监管模式

与美国强调发展创新不同，欧盟在人工智能治理上更突出风险防控与基本权利保护，形成以安全为导向的强监管模式。欧盟早在 2021 年就提出《人工智能法》草案，已于 2024 年 3 月正式通过该法，是世界范围内通过的首部人工智能综合立法。《人工智能法》奠定了欧盟的人工智能监管制度框架。其主要内容包括：第一，在监管对象上，该法将人工智能系统定义为“一种以机器为基础的系统，其设计运作具有不同程度的自主性，在部署后可表现出适应性，并且为了明确或隐含的目标，可从其接收的输入中推断出如何生成可影响物理或虚拟环境的预测、内容、建议或决定等输出”。第二，在监管方法上，主要采取基

[27] Biden J. Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. The White House[EB/OL]. (2023-10-30)[2026-03-09]. https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.

[28] The White House. FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI[EB/OL]. (2023-07-21) [2026-03-09]. https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/; The White House. FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Eight Additional Artificial Intelligence Companies to Manage the Risks Posed by AI[EB/OL]. (2023, September 12)[2026-03-09]. https://www.whitehouse.gov/briefing-room/statements-releases/2023/09/12/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-eight-additional-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/.

于风险的分级监管方法，区分禁止的人工智能实践、高风险人工智能系统、有限风险人工智能系统和最小风险人工智能系统，对不同风险等级的人工智能系统施加不同程度的监管措施，并专门针对通用人工智能作出规定。第三，在监管机构上，成立欧洲人工智能办公室，负责在欧盟层面对人工智能进行监管，推动《人工智能法》的实施。

在生成内容监管方面，欧盟首先特别强调透明性与可追溯性。例如，要求对深度伪造内容进行明确标识，防止误导公众；对大模型的训练数据来源进行合规说明，尤其是在版权保护领域强化数据使用合法性要求。其次，欧盟也更加强调事前合规。高风险系统在投放市场前需完成合格评估，并接受持续的市场监管。这种模式虽未在生成式人工智能领域全面实施许可制度，但通过风险评估、技术文件备案与监管机构审查等方式，强化了平台的事前责任。

欧盟所采取的强监管模式，正在助力其在人工智能监管方面获取更大的影响力和话语权，在人工智能监管领域形成“布鲁塞尔效应”。“布鲁塞尔效应”是指欧盟事实上通过市场机制将其法律推广至境外，导致其单边监管全球化的过程。欧盟在数字治理规则的建设上一直走在世界前列。早在《人工智能法》之前，欧盟就通过《通用数据保护条例》《非个人数据在欧盟境内自由流通框架条例》《开发数据指令》《数字内容治理》《数字市场法》《数字服务法》《数据法》等一系列重大立法，在数字治理领域产生了重要影响，形成了数字治理的“欧盟模式”。在人工智能治理领域，欧盟也在延续这一模式的诸多特点，与其他国家和地区特别是美国的治理模式形成了鲜明的对比。这种强监管模式主要采取对相关主体施加法定义务和责任的“硬法”方法，有国家强制力的保障，可以有效地督促相关主体履行义务和承担责任。但是如果施加的义务和责任过重，或者规则过于一刀切，容易对人工智能发展造成不利影响。

（三）中国：发展与安全并重的中间道路

在人工智能治理领域，中国探索形成了一条介于美国与欧盟之间的“中间道路”。中国的监管模式既不是美国式的高度依赖市场自律，也未完全采纳欧盟式的全面立法，而是在行政规章层面率先建立专门性规范，在鼓励技术创新与产业发展的同时，强化对特定人工智能风险的监管。

在生成式人工智能兴起的背景下，国家网信部门于 2022 年和 2023 年先后发布了《互联网信息服务深度合成管理规定》和《生成式人工智能服务管理暂行办法》，对向公众提供生成式人工智能服务的行为进行专门规范。在网络信息内容治理方面，这些规章主要有两个方面的发展。第一，在网络信息内容的生成环节，细化了相关主体防止生成违法和不良信息的义务。《深度合成管理规定》主要规定了深度合成服务提供者、深度合成服务技术支持者、深度合成服务使用者三类主体，《管理暂行办法》则主要区分了生成式人工智能服务提供者和生成式人工智能服务使用者两类主体。这些规章针对与网络信息内容生成相关的各个子环节均作出了细致的规定。以服务提供者的义务为例，这些规章中的规定覆盖了服务提供者在数据管理、内容管理、用户管理等各个子环节的义务。例如，在数据管理方面，服务提供者应当在训练数据选择过程中，采取有效措施防止产生歧视；应当依法开展预训练、优化训练等训练数据处理活动；进行数据标注的，应当制定标注规则，开展数据标注质量评估，对标注人员进行培训、监督和指导。在内容管理方面，服务提供者应当采取技术或者人工方式对输入和生成结果进行审核，发现违法和不良信息的，应当采取处置措施。在用户管理方面，服务提供者发现违法和不良信息的，应当对相关服务使用者采取处置措施。第二，在网络信息内容的传播环节，新增了相关

主体对人工智能生成内容进行标识的义务。《深度合成管理规定》第十六条至第十八条确立了服务提供者对人工智能生成内容进行标识的基本规则，包括：（1）标识的一般义务。服务提供者应当对其生成的内容添加不影响用户使用的标识。（2）显著标识的特别义务。如果服务提供者提供特定的深度合成服务，可能导致公众混淆或者误认的，应当在其生成内容的合理位置、区域进行显著标识。（3）标识的保护。任何组织和个人不得破坏对人工智能生成内容的标识。2025年3月，《人工智能生成合成内容标识办法》的发布，进一步完善我国的了人工智能生成内容标识义务制度。该办法围绕人工智能生成内容的标识问题作出集中规定，明确区分显式标识与隐式标识两种形式，并对服务提供者、用户等不同主体在内容生成、传播和平台管理环节中的标识义务作出具体规定。为配套该办法的实施，我国还制定了国家标准《网络安全技术 人工智能生成合成内容标识方法》（GB45438—2025），为如何添加标识提供了具体的指引。

综上，从美国发展导向的弱监管模式，到欧盟安全导向的强监管模式，再到中国发展与安全并重的制度探索，全球生成式人工智能内容监管呈现出多元路径并行的格局。不同模式反映出各法域在价值取向、法律传统与产业结构上的差异，也为未来国际协调与规则趋同提供了基础。

四、完善人工智能生成内容治理的建议

（一）内容识别：技术与制度协同发展

面对人工智能生成内容的识别挑战，有必要在技术与制度两个层面协同发展，进一步完善生产内容的治理机制。

第一，应强化生成内容识别技术体系建设。一方面，要推动识别技术研究发展，支持相关机构研究生成内容识别技术，特别是要加强对生成内容识别技术难点的攻关，包括：强化对生成文本识别技术的研究，聚焦于提升生成文本识别技术的准确率，探索鲁棒性更强、适应短文本的生成文本标识方法，提高事后检测的生成文本识别技术的泛化性；持续跟踪识别技术对抗技术的发展，研究如何增强识别技术的鲁棒性，以防范对生成内容标识的破坏以及对检测技术的规避。另一方面，要完善识别技术的配套服务机制，建立公共平台，方便用户查询查询已知的带水印输出的模型及其识别服务，或者由平台提供不同模型的统一识别服务；倡议在国际层面制定识别技术和标识的标准以及建立相应的组织，扩大我国在标准制定和国际组织中的话语权和影响力。

第二，应完善生成内容信息披露义务制度。一方面，要细化生成内容标识义务的规则。区分不同的场景，考虑内容标识难度、用户体验影响、法律风险大小等因素，通过部门规章或行业标准进一步细化服务提供者在不同场景下标识义务的具体要求。明确错误识别的处理规则，赋予识别对象的利益相关方对识别结果提出异议的权利，细化异议处理的程序规则，如双方未能达成一致，可由利益相关方提起诉讼，由法院来判决，识别方根据法院认定结果予以处理，应明确错误识别的责任，如识别方存在过错，还应就识别错误对利益相关方造成的损害承担赔偿损失责任。另一方面，在现有标识义务制度的基础上，需要进一步构建全面的生成内容信息披露义务制度。信息披露义务的内容应从“是否使用人工智能”扩展至“如何使用人工智能”，方式应从单一的标识义务拓宽至标识和说明双重义务，并基于信息成本对不同披露主体的义务进行合理配置。

（二）平台治理：新型平台的规则构建

生成式人工智能服务提供者等新型平台的出现，为更新平台规制规则提供了契机，有必要在制度层面探索新型平台的规制规则。

第一，应明确生成式人工智能服务提供者的法律定位。在现有网络服务提供者类型基础上，增加“生成式人工智能服务提供者”这一独立类别，将其与传统的网络服务提供者类型区分开来。通过制定专门性的立法或规章，对生成式人工智能服务提供者的法律规制规则坐出具体规定。

第二，应根据服务提供者的信息管理能进行分类规制。例如，生成式人工智能服务提供者依托的模型来源有所不同，提供服务的对象也不尽相同，这会影响其信息管理能力。根据服务提供者依托的模型来源，可以将服务提供者分为基于自研模型的服务提供者、基于二次训练模型的服务提供者和基于第三方模型的服务提供者。这三类服务提供者在对训练数据和模型的管理能力上存在较大差异。基于自研模型的服务提供者对训练数据和模型有较强的管理能力，因而在训练数据选择和模型训练方面可能负有较高的注意义务。基于二次训练模型的服务提供者一般仅对二次训练过程有较强的管理能力，因而主要对二次训练过程负有相应的注意义务。基于第三方模型的服务提供者对训练数据和模型的管理能力最弱，通常没有参与模型的训练，无法采取与训练有关的预防和处置措施。根据服务提供者提供服务的对象，也可以将服务提供者分为面向应用开发者的服务提供者和面向终端用户的服务提供者。这两类服务提供者在对用户行为的管理能力上存在较大差异。向应用开发者提供应用程序编程接口（API）的服务提供者通常难以监控应用程序中的终端用户行为。相比之下，直接向终端用户提供服务的提供者商更有能力直接对用户进行监督和管理，因此在这方面会负有更高的注意义务，理应采取更多的措施来预防和处置用户的侵害行为。

第三，应完善平台举证责任与信息披露规则。由于生成式人工智能具有黑箱性，生成式人工智能服务

提供者应承担一定程度的举证责任与信息披露义务，包括披露记录生成日志、保存提示语与输出内容，以便在争议发生时进行责任认定。若平台拒绝提供必要信息，可依法推定其存在过错。

（三）责任分配：内容损害的责任规则

第一，在多重主体之间分配责任时，需要结合具体情形，考虑不同主体行为对生成侵权内容的原因力，来确定责任承担的主体。例如，如果生成侵权内容的原因是服务提供者在训练数据中纳入了相关侵权内容且未能采取有效的预防措施，导致即使用户仅输入相关性较低的提示词，人工智能也会生成侵权内容，那么侵害行为的主体应当是服务提供者。相反，如果生成侵权内容主要是由于用户输入中包含了相关侵权内容，使人工智能在输出时记忆并重现了输入中的内容，那么侵害行为的主体显然应当是用户。此外，如果服务提供者未能采取有效的防范措施，且用户输入的提示词中也包含侵权内容，生成内容是由服务提供者与用户的行为共同作用所致时，则可以认为服务提供者和用户构成共同侵权。

第二，考虑到责任分配可能涉及不同的场景，可以遵循场景化的原则，考虑不同场景下具体对策在预防风险方面的成本与收益，选择社会效率最优的责任分配对策。例如，在中国，法院时常会考虑各种因素来确定服务提供者在具体场景下是否存在过错。我国的司法解释曾就网络服务提供者过错的认定作出规定，列举了确定过错时应当考虑的多种因素，以指导法院在具体案件中作出判断。虽然我国现行法律法规和司法解释尚未对生成式人工智能服务提供者的过错认定作出类似的规定，但在司法实践中，已有法院参照网络服务提供者过错认定的方法进行分析。例如，在杭州奥特曼案中，法院就认为：“对于过错的认定规则，应综合考量生成式人工智能服务的性质、当前人工智能技术的发展水平、避免损害的替代设计的可行性与成本、可以采取的必要措施及其效果、侵权责任的承担对行业的影响等因素，通过动态地调整过错的认定标准，将平台注意义务控制在合理的程度。”[29] 可见，网络服务提供者过错认定的方法为生成式人工智能服务提供者过错的认定提供了借鉴的经验，方便法院考虑不同的场景来做出合理的认定。

[29] 浙江省杭州市中级人民法院（2024）浙 01 民终 10332 号民事判决书。