国家标准《网络安全技术　人工智能生成合成内容标识方法》（征求意见稿）编制说明

一、工作简况

（一）任务来源

根据国家标准化管理委员会2024年下达的国家标准制修订计划，强制性国家标准《网络安全技术　人工智能生成合成内容标识方法》由中央网络安全和信息化委员会办公室提出，委托全国网络安全标准化技术委员会执行，主要起草单位为中国电子技术标准化研究院，计划号：20241842-Q-252。

（二）制定背景

生成式人工智能已成为继移动互联网技术之后最大的一波技术浪潮，同时也带来了新的安全风险和挑战。随着人工智能技术的发展，人工智能生成合成内容日益逼真，网络传播内容是否由人工智能生成合成难以分辨，社会上已出现多起利用生成合成内容传播虚假新闻、引发社会舆情，或是利用生成合成内容进行诈骗的案件。人工智能生成合成内容存在被误用、滥用、恶意使用的安全风险，严重影响国家安全，危害广大人民群众在网络空间的合法权益。

2022年11月，国家网信办等三部门发布《互联网信息服务深度合成管理规定》（以下简称“《规定》”），提出了深度合成服务提供者对使用其服务生成或者编辑的信息内容进行标识的要求。2023年7月，国家网信办等七部门发布《生成式人工智能服务管理暂行办法》（以下简称“《办法》”），要求生成式人工智能服务提供者对图片、视频等生成内容进行标识。

为落实《规定》《办法》相关要求，切实维护国家安全和公共利益，制定本标准，对人工智能生成合成内容的标识方法提出规范。本标准对防范人工智能生成合成内容引发安全风险、提升人工智能安全水平起到规范作用，促进人工智能行业安全发展。

（三）起草过程

1、预研阶段

（1）2023年11月，组建编制组，编制形成第一版草案。

（2）2023年11月-2024年2月，标准编制组开展广泛调研，并多次组织组内研讨，持续完善标准草案，对标准草案进行多轮迭代。

（3）2024 年 3 月，组织 10 余家相关企业多次开展调研、召开研讨会，收集企业反馈材料，进一步完善标准草案。

（4）2024 年 4-5 月，继续联系重点企业，征求企业意见，结合企业反馈迭代标准草案版本，形成标准可行性研究报告。

2、起草阶段

（1）2024 年 6 月 25 日，国家标准化管理委员会下达本强制性国家标准的制定计划，标准正式立项。

（2）2024 年 7 月 3 日，成立强制标准工作专班，由中国电子技术标准化研究院牵头，国家计算机网络应急技术处理协调中心、浙江大学、中国科学院软件研究所等共同组成。

（3）2024 年 7 月 3 日-15 日，标准工作专班组织开展基础调研，对标准内容进行研讨，分工修改完善标准草案。

（4）2024 年 7 月 16 日-30 日，先后组织 20 余家重点通用类与垂域类企业召开 3 次研讨会，征求企业意见与反馈，完善标准草案。

（5）2024 年 8 月 9 日，组织召开专家研讨会，征求专家意见建议，并修改完善标准草案。

（6）2024 年 8 月 10 日-25 日，根据专家意见对标准草案多轮修改与完善，更新标准草案文本。

（7）2024 年 8 月 26 日，再次联系 20 余家重点通用类与垂域类企业召开研讨会，征求企业意见与反馈，完善标准草案。

（8）2024 年 8 月 30 日，组织专家评审会，专家一致同意该项标准通过评审，根据专家意见修改完善后形成征求意见稿。

二、编制原则、强制性国家标准主要技术要求的依据及理由

（一）标准编制原则

本标准的编制原则是：

1）通用性：生成合成服务提供者和内容传播服务提供者均可依据本标准开展对人工智能生成合成内容进行标识活动。

2）可行性：确保标准中技术要求可验证、可操作。为此，本标准编制过程中与科研机构、相关企业、专家进行了多轮研讨。

3）符合性：符合《规定》、《办法》等国家有关法律法规和已有标准规范对于标识的相关要求。

（二）主要技术要求及其确定依据

本标准给出了人工智能生成合成内容标识方法。本标准适用于规范生成合成服务提供者和内容传播服务提供者对人工智能生成合成内容开展的标识活动。

本标准的主要技术要求包括：显式标识方法，隐式标识方法。

显式标识方法包括文本内容显式标识、图片内容显式标识、音频内容显式标识、视频内容显式标识和交互场景界面显式标识。规范了在不同模态的内容上及交互场景界面上进行显式标识的方法，依据为《规定》第十七条“深度合成服务提供者提供以下深度合成服务，可能导致公众混淆或者误认的，应当在生成或者编辑的信息内容的合理位置、区域进行**显著标识**，向公众提示深度合成情况”。对目前已对公众提供人工智能生成合成内容服务的产品进行了调研，发现内容显式标识与交互场景界面显式标识已有多家重点企业进行了实践。

隐式标识方法包括元数据隐式标识。规范了在提供文件形式的生成合成内容时，添加元数据隐式标识的方法，依据为《规定》第十六条“深度合成服务提供者对使用其服务生成或者编辑的信息内容，应当采取技术措施添加**不影响用户使用的标识**，并依照法律、行政法规和国家有关规定保存日志信息”。调研了国内国际不同格式文件元数据写入技术，确保主流文件格式的元数据可写入、可识别。

为落实《规定》《办法》相关要求，切实维护国家安全和公共利益，在有关主管部门的指导下，编制组广泛调研国内外生成式人工智能技术研发机构及企业所开展的探索和应用，编制组成员分工合作，完成了技术可行性与法律法规符合性的调研分析工作，确保了标准的可落地、可实施。

（三）修订前后技术内容的对比[仅适用于国家标准修订项目]

不涉及。

三、与有关法律、行政法规和其他强制性标准的关系

《规定》第十六条提出，深度合成服务提供者对使用其服务生成或者编辑的信息内容，应当采取技术措施添加不影响用户使用的标识，并依照法律、行政法规和国家有关规定保存日志信息。《规定》第十七条提出，深度合成服务提供者提供以下深度合成服务，可能导致公众混淆或者误认的，应当在生成或者编辑的

信息内容的合理位置、区域进行显著标识，向公众提示深度合成情况：

（一）智能对话、智能写作等模拟自然人进行文本的生成或者编辑服务；

（二）合成人声、仿声等语音生成或者显著改变个人身份特征的编辑服务；

（三）人脸生成、人脸替换、人脸操控、姿态操控等人物图像、视频生成或者显著改变个人身份特征的编辑服务；

（四）沉浸式拟真场景等生成或者编辑服务；

（五）其他具有生成或者显著改变信息内容功能的服务。

深度合成服务提供者提供前款规定之外的深度合成服务的，应当提供显著标识功能，并提示深度合成服务使用者可以进行显著标识。

《办法》第十二条提出，提供者应当按照《规定》对图片、视频等生成内容进行标识。

本标准依据《规定》第十六、十七条，将标识分为显式标识与隐式标识，并分别规范了显式标识与隐式标识的标识方法，对《办法》《规定》中的要求进行细化，本标准与现行法律、法规以及国家标准不存在冲突与矛盾，与其他标准配套衔接。

四、与国际标准化组织、其他国家或者地区有关法律法规和标准的比对分析

欧盟《人工智能法》和《数字服务法》两部法律涉及 AI 标识，根据这两部法律的规定，目前主流的生成式人工智能属于有限风险人工智能系统，须承担标识义务，标识技术包括水印、元数据识别、指纹识别、加密方法、日志记录等。美国拜登政府颁布的《关于 AI 的安全和可信赖开发和使用的行政命令》为美国 AI 标识技术标准、指南等文件的出台和工作的开展提供指引。新加坡《生成式人工智能治理模型框架》、加拿大《生成式人工智能行为守则》均提及内容标识技术的应用。国际组织层面，七国集团通过《开发高级人工智能的组织的国际指导原则》，呼吁采用水印等技术使用户能够识别人工智能生成内容。

目前，国际标准化组织 ISO/IEC JTC 1/SC 42（人工智能分委员会）和 SC 27（信息安全、网络空间安全和隐私保护分委员）分别收到了来自加拿大和我国的关于 AI 标识的标准化贡献，但都还未正式立项。NIST 发布的标准《数字内容透明度技术方法概述》中，主要内容有验证内容并跟踪其来源、标记合成内容、检测合成内容等。标准第 3 节提到标记合成内容分为内容标签、可见水印、披露字

段等直接向用户披露内容创作过程中使用 AI 情况的技术；与隐形水印、数字指纹、嵌入的元数据等间接披露技术。本标准未规定具体的技术指标。C2PA 发布的《C2PA 技术规范》中，主要是采用可追溯的元数据来保证内容的真实性。

本标准与现有国际政策、标准相比，明确了标识的形式，且技术完备，适用性强。本标准的提出可以引领人工智能技术的规范、安全发展，为我国的 AI 标识国际标准贡献提供支撑。本标准与现有国际政策、标准不冲突，本标准的颁布实施对国际贸易不会带来壁垒性的影响。

五、重大分歧意见的处理过程、处理意见及其依据

本标准修订过程中无重大分歧。

六、对强制性国家标准自发布日期至实施日期之间的过渡期（以下简称过渡期）的建议及理由

考虑到执行人工智能标识生成与检测技术复杂性相对较低，技术改造所需时间较短，并且由于已有大量制作平台已经开展部分内容标识工作、预计具备短期内适应标注的能力。建议本标准自发布之日起，过渡期为半年。半年以后，正式实施。

七、与实施强制性国家标准有关的政策措施

本标准是《人工智能生成合成内容标识办法》的配套强制性国家标准，该文件当前为征求意见稿。文件第十一条中规定“服务提供者应当按照有关强制性国家标准的要求进行标识”。其中所指“有关强制性国家标准”即为本标准。

本标准编制过程中配套建设人工智能生成合成内容标识验证平台，对后续标准的实施起到支撑作用。同时计划后续对重点企业进行宣讲、培训工作，开展相应的检测、认证服务。

八、是否需要对外通报的建议及理由

本标准既是对我国的《规定》和《办法》的标准支撑，同时也可响应国际标准化组织、其他国家或者地区有关 AI 标识的法律法规和标准。本标准与现有国际政策、标准相比，明确了标识的形式，且技术完备，适用性强。本标准的提出可以引领人工智能技术的规范、安全发展，为我国的 AI 标识国际标准贡献提供支撑。

另外，本标准为强制性国家标准，还影响到国际人工智能企业在我国境内提

供生成合成内容服务，故建议对外通报。

九、废止现行有关标准的建议

不涉及。

十、涉及专利的有关说明

本标准不涉及相关专利、知识产权、著作权等内容。

十一、强制性国家标准所涉及的产品、过程或者服务目录

本标准适用于生成合成服务提供者和内容传播服务提供者对人工智能生成合成内容开展的标识活动。以及支撑标识活动的相应产品和服务开发和应用。

十二、其他应当予以说明的事项

无。

标准编制组

2024 年 9 月 13 日