AI Compliance Frontline: A Deep Dive into AIGC Labeling

Author: YANG, Jianyuan ZHAO, Yaze

2025年3月14日，国家网信办等四部门联合发布《人工智能生成合成内容标识办法》（“《标识办法》”），《标识办法》与其配套的强制性国家标准《网络安全技术 人工智能生成合成内容标识方法》（“《标识方法》”）均将自2025年9月1日起施行。

这意味着，AIGC标识义务作为我国人工智能专项监管的独特一隅，已率先走向成熟与落地。其原因既在于AIGC虚假信息对网络生态信任根基的动摇，误导诱骗、判断依赖、名誉损害、舆论污染等风险已然显现，亟需有所规范；又在于其已经具备相对成熟的解决方案，相较于从根本上解决存在技术卡点的模型幻觉等问题，AIGC标识显然更容易实现。

根据《标识办法》，我国AIGC标识要求具有广泛的辐射面，并不仅仅局限于内容生成合成环节这一起点，其后续的分发、传播、甚至使用环节均被涵盖在内。随着AIGC服务深度融入现代生产生活，除大模型企业外，各行业企业均可能涉及AI生成合成内容的业务内嵌与外传，进而可能需要履行标识义务，包括但不限于标识添加、检测、追踪、管理、提示等。

本文将以国内AIGC标识新规出台为切入点，重点解答两个问题：（1）AI生成合成内容“生产-分发-传播-使用”链条中的各方需履行何等标识义务；（2）AIGC标识合规方案能否全球通用。

一、中国智慧：全链留痕，欲登高则脚踏实地逐步而行之

我国AIGC标识监管已然形成较为完整的监管体系，以《标识办法》和《标识方法》作为集大成者，致力于在规范AI生成合成内容衍生流转的同时，提供可落地的具体合规指引，最大限度地保护行业创新与发展。

文件名称	发布时间	发文机关	标识相关内容
《网络音视频信息服务管理规定》	2019 年 11 月	国家网信办、文旅部、国家广电总局	第十一条第一款　网络音视频信息服务提供者和网络音视频信息服务使用者利用基于深度学习、虚拟现实等的新技术新应用制作、发布、传播非真实音视频信息的，应当以显著方式予以标识。
《互联网信息服务算法推荐管理规定》	2021 年 12 月	国家网信办、工信部、公安部、国家市监局	第九条第一款　算法推荐服务提供者应当加强信息安全管理，建立健全用于识别违法和不良信息的特征库，完善入库标准、规则和程序。发现未作显著标识的算法生成合成信息的，应当作出显著标识后，方可继续传输。
《互联网信息服务深度合成管理规定》	2022 年 11 月	国家网信办、工信部、公安部	第十七条第一款　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。
《生成式人工智能服务管理暂行办法》	2023 年 7 月	国家网信办、国家发改委等七部门	第十二条　提供者应当按照《互联网信息服务深度合成管理规定》对图片、视频等生成内容进行标识。
《关于加强“自媒体”管理的通知》	2023 年 7 月	国家网信办	第三条　使用技术生成的图片、视频的，需明确标注系技术生成。
《网络安全标准实践指南——生成式人工智能服务内容标识方法》	2023 年 8 月	国家信安标委（TC260）	围绕文本、图片、音频、视频四类生成内容给出了内容标识方法，可用于指导生成式人工智能服务提供者提高安全管理水平。
《人工智能生成合成内容标识办法》	2025 年 3 月	国家网信办、工信部、公安部、国家广电总局	围绕人工智能技术生成、合成的文本、图片、音频、视频、虚拟场景等生成内容对内容生产、分发、传播、使用链条上多方主体规定全链路标识义务。
《网络安全技术　人工智能生成合成内容标识方法》	2025 年 3 月	国家市监局、国家标准化管理委员会	为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标识活动，制定人工智能生成合成内容的显式标识和隐式标识方法标准。

▲ 表1：中国标识相关主要立法

（一）标识办法：贯穿全生命周期、平衡发展与安全的义务体系

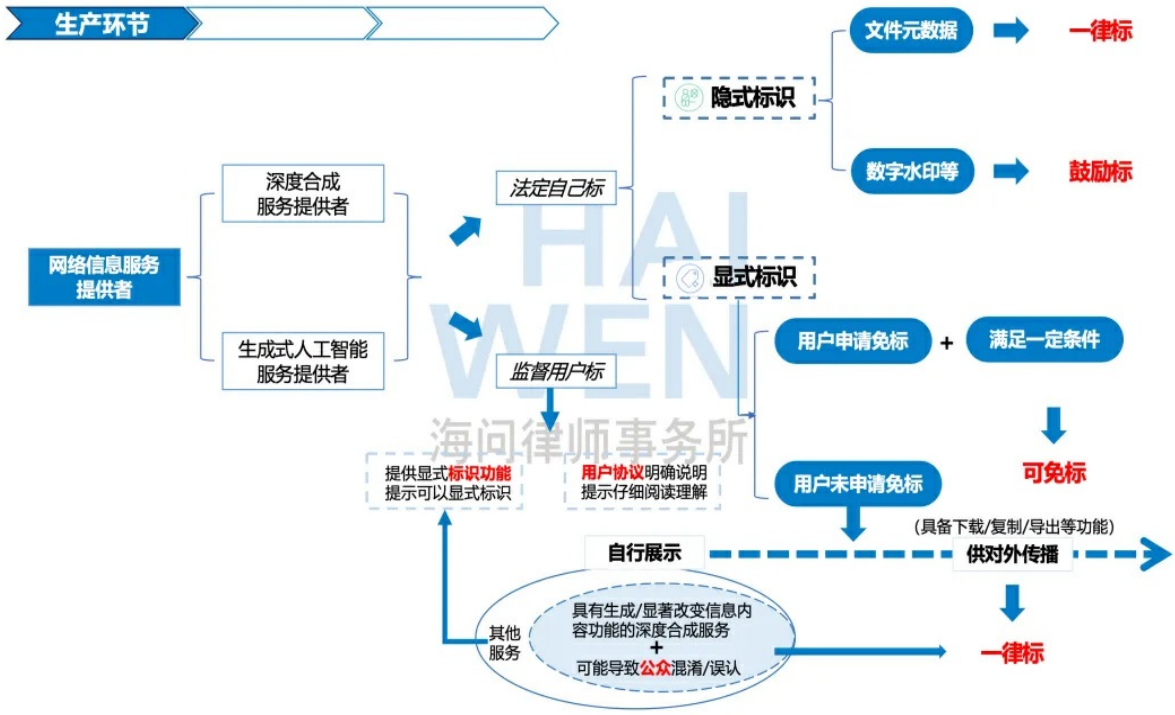
《标识办法》规范AI生成合成内容“生产-分发-传播-使用”的全生命周期，义务主体覆盖提供生成合成服务的网络信息服务提供者、互联网应用程序分发平台、互联网应用程序服务提供者、提供网络信息内容传播服务的服务提供者，以及用户。

就法律责任性质而言，违反标识义务可能导致行政处罚和违约责任。《标识办法》未直接设立罚则，而是将其转引至“有关法律法规及部门规章”。对于提供者、传播者与分发平台，其可能因违反标识义务而依相关立法受到行政处罚，或视其合同约定承担违约责任。而对于用户，暂未见相关立法对用户的标识义务设立具体罚则，其未履行标识义务主要需承担违约责任。

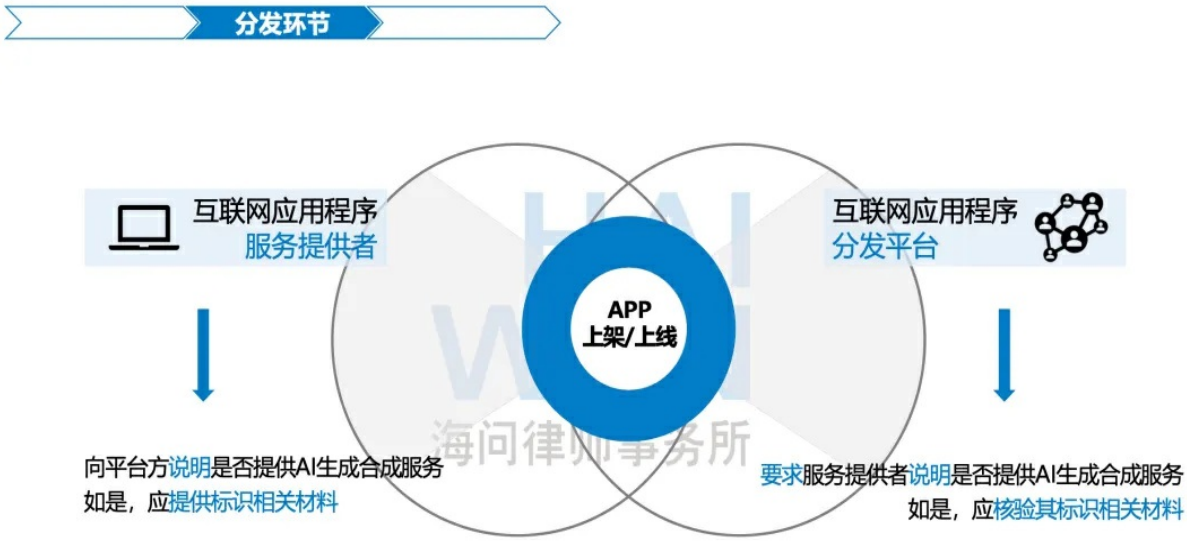
《标识办法》体现了发展与安全相平衡的监管思路。具体而言：

- 一方面，《标识办法》立足企业技术难点与合规成本考量，在安全基线之上为企业合理减负。例如，就隐式标识而言，《标识办法》就较易实现、成本较低的元数据标识作出强制添加要求，而对于可能仍存技术卡点、成本相对较高的数字水印，则仅提出鼓励性添加倡议。

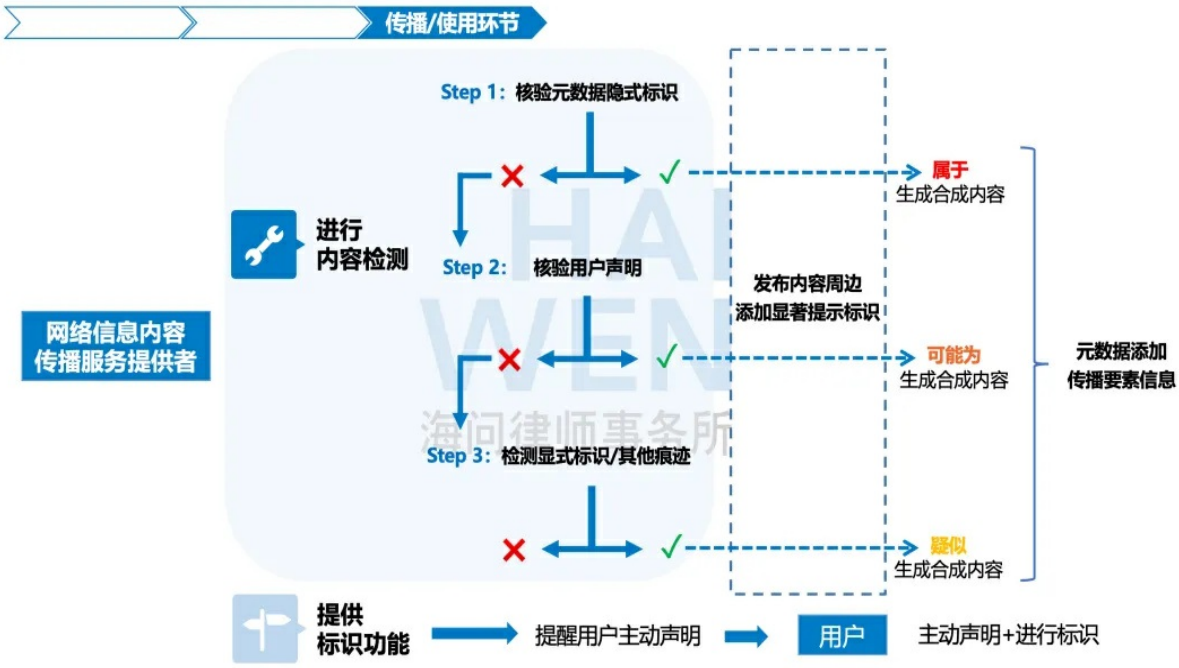
- 另一方面，《标识办法》容留弹性空间以满足实践发展需求。例如，考虑到特定场景下显式标识可能会影响AI生成合成内容的正常使用，《标识办法》允许用户在受用户协议约束的前提下申请使用未添加显式标识的AI生成合成内容，既顾及实践中多元化的场景使用需求，同时也一定程度上为服务提供者提供了“避风港”（前提是明确用户协议、依法留存日志）。



▲图1：内容生产环节标识义务分析



▲图2：内容分发环节标识义务分析



▲ 图3：内容传播/使用环节标识义务分析

（二）标识方法：实操性强、进阶式的具体标识方案

显式标识关乎AI生成合成内容的透明度，而隐式标识意在实现溯源监督。

- 从特性上来看，两者的核心差异在于是否可被感知。显式标识可被用户明显感知，隐式标识则不易被用户明显感知。

- 基于该等特性，两者具有差异化的功能。显式标识基于其显著性得以对公众起到直接的提示效果，可协助公众快速辨别AI生成合成内容，故而其标识内容以“AI生成/合成”作为核心要素，功能亦在于显著提示；隐式标识基于其隐蔽性、鲁棒性得以对内容来源与流转实现较为完整的记录，故而其标识内容亦以记录内容属性和流传信息为主，功能即在于溯源监督。

《标识方法》系《标识办法》的配套性文件，为企业履行标识义务提供了详细具体、可落地性强的标识指引。考虑到关于显式标识的规定清晰、直观易懂，我们在此不再作赘述；关于隐式标识的原理及内在逻辑介绍较少，我们在此简要分享关于元数据、数字水印的基本原理。

《标识办法》与《标识方法》提出了进阶式的隐式标识方案。

1. 基础性强制义务：文件元数据标识

文件元数据是指按照特定编码格式嵌入到文件头部的描述性信息，用于记录文件来源、属性、用途等信息内容，通俗而言，即文件的“说明书”。例如，针对一张照片，其元数据会描述照片的拍摄地点、时间、拍摄设备、存储大小与路径、文件类型等详细信息。在大多数情况下，文件创建时会自动产生一些基础性的元数据信息（如文件创建时间、修改时间、文件大小等），用户亦可在后续手动添加或更新元数据。

关于文件元数据隐式标识，通常而言是将需要嵌入的标识信息添加在文件现有元数据中。《标识方法》对其标识内容作出了具体规定，包括：1）生成合成标签要素（即区分“属于”/“可能为”/“疑似为”生成合成内容）；2）生成合成服务提供者要素（即提供者的名称或编码）、内容制作编号要素（即提供者对该内容的唯一编号）；3）内容传播服务提供者要素（即传播者的名称或编码）、内容传播编号要素（即传播者对该内容的唯一编号）。

总体而言，文件元数据隐式标识具有易实现、标识成本低的优点，但同时也存在易擦除和篡改的缺陷。

2. 进阶性倡导建议：数字水印标识等内容隐式标识

数字水印是指通过特定算法与规则而嵌入到数字内容（与元数据嵌入到文件头部有所差异）中的标记信息。数字水印的相对概念是更为原始的“物理水印”，典例如钱币上的水印，同样得以起到溯源、鉴伪的作用。

数字水印技术随着数字媒体技术的发展而兴起，此前存在诸如以下的应用场景，但整体应用尚不普遍：1）版权保护：虽然不能起到直接的防侵权作用，但得以协助确认版权归属、作为版权侵权的证据，目前常被用于强版权意义场景，如图片版权、模型版权保护（以预防模型蒸馏攻击）等。2）用户提示：例如在数据中添加用户信息以警示用户依法使用内容，同时帮助企业进行免责。此外，Google此前率先针对AI生成合成内容推出SynthID水印工具，引发业内的广泛关注与宣传，亦为其树立负责任AI企业的良好社会形象有所助力。

考察数字水印的关键性指标通常包括：1）隐蔽性：嵌入水印后不易被察觉，不影响原始内容的表达；2）鲁棒性：对数据内容作正常加工处理，水印仍得以被完整准确提取；3）安全性：在受到攻击的情况下，水印难以被破坏或篡改。此外还包括容量性（即可嵌入的水印数据量）、实时性（即随着原始内容的产生而实时嵌入水印，例如直播视频水印的嵌入）等指标。

以上指标存在制约甚至互斥关系，由此带来一定的技术难题。我国标识监管充分考量行业技术难点与合规成本，仅将数字水印作为倡导性标识方案。在官方答记者问中，明确提出不对“在文本内容中添加隐式标识”“在多媒体文件中添加数字水印”作强制要求，试解析原因如下。

- 就文本水印而言，其相较于图片、音视频等带有颜色、结构、频率、时间等要素的内容呈现形式，文本内容以字符表达、信息密度更高，导致在文本内容中嵌入少量信息即可能影响原有文本语义。故而，目前仅能在模型底层服务而非应用层中添加水印，但在模型层添加水印又可能会影响模型生成内容质量。

- 就多媒体文件而言，其载体内容丰富复杂，且实践中多媒体处理方式与标准在载体属性、压缩标准、传输信道等诸多方面均快速迭代，导致平衡水印的隐蔽性和鲁棒性颇具挑战。

二、全球监管：求同存异，然境内外标准化合规之路漫漫

对AI生成合成内容设立标识要求已成为主流法域的监管通识。与欧盟、美国等刚确立上层立法要求的情况相比较，中国的体系化AI标识监管已先一步走向落地。

- 中国：以《互联网信息服务算法推荐管理规定》《互联网信息服务深度合成管理规定》《生成式人工智能服务管理暂行办法》作为立法基底，以《标识办法》作为专项立法要求，以《标识方法》作为标识具体落地方案，后两者将于2025年9月1日起实施。

除已于2025年3月14日发布的《网络安全标准实践指南——人工智能生成合成内容标识 服务提供者编码规则》外，相关元数据标识规范、各应用场景的标识方法等一系列推荐性标准、实践指南亦将逐步推出。

- 欧盟：欧盟《人工智能法》第50条对特定AI系统的透明度要求作出规定，以此作为标识义务的主要依据。《人工智能法》已于2024年8月颁布生效，但关于标识义务的要求将于2026年8月2日起才会正式实施。

- 美国：美国在联邦层面尚未正式出台关于标识的专项立法，但各州已开始探索AI标识规制，典例如《加州人工智能透明度法案》（California AI Transparency Act，“CAITA”），该法已经加州州长签署通过，将于2026年1月1日起生效。

国家	文件名称	标识相关内容
新加坡	《生成式人工智能治理模型框架》	探索**数字水印**和**加密溯源**两种技术方案。
韩国	《信息通信网络利用促进及信息保护等相关法律》	科学信息通信部长或通信委员会应**开发和分发技术**，以识别通过信息和通信网络传播的信息中使用 AI 技术创建的声音、图像或视频等**虚假信息**。
日本	《人工智能业务指南》	在技术可行的情况下，开发并部署可靠的**内容认证和来源机制**，例如**水印**或其他技术，以使用户能够识别由 AI 生成的内容。努力开发工具和应用程序编程接口（**API**），让 AI 业务用户和非业务用户能够通过**水印**确定特定内容是否由高级 AI 系统生成。建议引入其他机制（比如**声明标签**），以帮助使用 AI 的业务用户和非业务用户知晓他们正在与 AI 系统进行交互。
越南	《数字技术产业法草案》	要求企业以**机器可读**和**可检测**的**格式标记** AI 系统的输出。虽然有一些方法可以实现这一目标，如在文件中添加元数据或在输出内容本身中添加水印，但用户可以轻松删除这些信号，从而限制了其有效性。
沙特阿拉伯	《面向公众的生成式人工智能指南》	在 AI 透明度和可解释性的可采取措施中，列举了使用**水印**等工具来帮助识别 AI 生成的内容。

▲表2：其他法域标识监管相关进展

就标识义务主体而言，各主流法域均以“生成合成服务提供者”（各法域下概念可能略有不同）作为核心义务主体，而在各相关方是否负有标识义务方面存在监管差异。

- 中国：依据AI生成合成内容生命周期进行全链路监管，义务主体覆盖提供生成合成服务的网络信息服务提供者、互联网应用程序分发平台、互联网应用程序服务提供者、提供网络信息内容传播服务的服务提供者以及用户。

- 欧盟：形成以“特定AI系统提供者+部署者”为核心的双轨制义务主体体系。考虑到欧盟《人工智能法》之大一统属性，其覆盖范围包括但并不局限于生成式AI。具体包括：

  - 两类提供者：与自然人进行直接交互的AI系统提供者、生成合成类AI系统（含通用目的AI系统）提供者；

  - 四类部署者：情感识别系统或生物特征分类系统的部署者、生成或操纵深度伪造内容（含图像和音视频）的AI系统部署者、生成或篡改公益相关公开文本信息的AI系统部署者。

- 美国加州：每月拥有超过100万名访客或用户，并且在加州地理范围内对外公开可用的生成式AI系统提供者。以访客或用户数量作为门槛划定“受管辖的提供者”（Covered Provider），再次体现了美国立法中小企业豁免的监管思路。

就具体标识义务而言，各主流法域的标识思路具有相似性，实质上均涉及显式标识和隐式标识，但在标识范围及标识检测等方面存在监管差异。

- 中国：明确划分显式标识和隐式标识，如上文所分析，显式标识要求的强制性受环节及内容类型等因素影响而有所区分，隐式标识则原则上均需强制添加。

- 欧盟：未明确设有显式标识和隐式标识的概念，但实质要求比较近似：

  - 《人工智能法》要求前述“两类提供者”及“四类部署者”最迟在系统与自然人首次互动或接触时提供清晰可辨（clear and distinguishable）的告知信息，与显式标识概念有所接近；

  - 针对生成合成类AI系统提供者，除不实质改变输入内容及执法场景等法定例外情形外，立法强制要求其以机器可读（Machine-Readable）且可检测的方式对内容进行标注，如水印、元数据标识、指纹等，与隐式标识的监管思路较为相像。

- 美国加州：明确划分显式披露（Manifest Disclosure）和隐式披露（Latent Disclosure），受管辖的提供者应：

  - 允许用户选择添加显式披露，以揭示相关内容（含图像和音视频，下同）为AI生成内容；

  - 在内容中包含隐式披露，在技术可行范围内传达提供者名称、系统名称和版本号、内容创建时间以及唯一标识符等信息；

  - 此外，CAITA明确要求受管辖的提供者应免费向用户提供AI检测工具，使后者得以用于评估相关内容是否由前者的生成式AI系统创建或修改。本项要求为CAITA的创新性要求，中国和欧盟在立法层面则尚未见此类明确规定。

不过，尽管各主流法域的监管思路有相似之处，但就具体技术实现方面，目前国际上并未达成统一的标识标准，而囿于地缘政治，未来的标准互认也仅能谨慎乐观。这意味着，企业在后续出海过程中，可能不得不针对性地推出当地AIGC标识合规方案，一定程度上会增加技术研发和法律合规成本。

- 2025年1月23日，美国总统特朗普签发名为《破除人工智能领域阻碍美国领先地位的壁垒》（Removing Barriers to American Leadership in Artificial Intelligence）的行政令，美国人工智能政策“去监管化”的导向毫无疑问，但致力于形成美国主导的AI安全标准（AI模型检测、红队测试等）对美国在全球AI领域形成绝对主导权同等重要。

- 具体到AIGC标识领域，2021年2月，Adobe领导的内容真实性倡议（CAI）和微软与BBC领导的“原点项目”（Project Origin）联合起来，形成了内容来源与真实性联盟（C2PA），其使命即在于建立内容真实性相关技术标准。目前，该联盟与ISO/IEC等国际标准组织正积极推动建立跨国界的标识技术标准体系，旨在构建一个具有全球互操作性的内容真实性认证框架。

2025-03-19