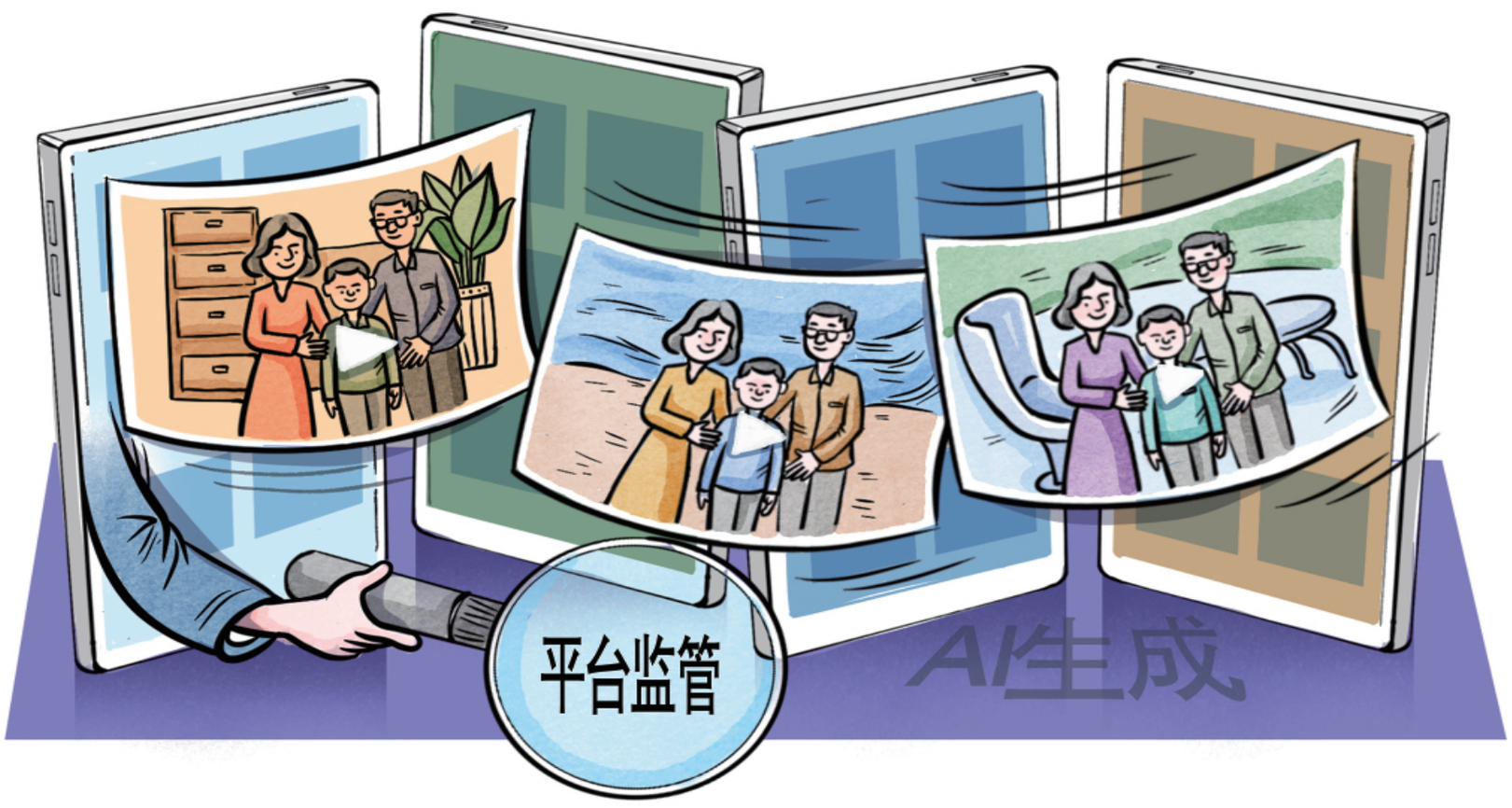
平台管理效果参差不齐 “去水印”即可绕过监管

AI时代的真伪博弈



□ 本报记者 赵 丽
□ 本报实习生 王艺霏

近日，安徽一名女子使用AI工具生成了高度逼真的“流浪汉卧坐餐厅”等图片，发送给在外聚餐的丈夫，本想测试其反应，不料丈夫信以为真，当即报警求助。

民警现场核实情况，确认是虚假信息，对该女子的行为进行了提醒：“AI技术生成的内容虽逼真，但以此制造恐慌并引发警方出警，属于浪费公共资源的行为。”

类似由AI生成图片引发的“乌龙事件”并非个例：广东广州某小区业主的孩子用AI生成图片后发给父母称“流浪汉试图闯进家门”，图片被转到业主群后引起楼栋业主恐慌；近日，多名电商平台卖家曝光，有消费者利用AI工具生成商品变质、损坏的图片，以此申请“仅退款”；山东威海网民张某为博取关注，使用AI工具编造“在监狱直播带货”图片的谣言信息并发布，引发大量网民关注讨论，张某因扰乱公共秩序已被公安机关处以行政处罚。

今年9月1日，《人工智能生成合成内容标识办法》正式实施，要求人工智能生成合成内容应当通过显著标识、数字水印等方式进行显式与隐式标注，以保障内容的可识别性。这标志着，AI生成内容全面进入“持证上岗”时代。

然而，《法治日报》记者近日对多个主流平台展开的调查显示，新规落地效果参差不齐，仍有大量未标识内容在各大平台持续传播。这些愈发逼真的AI生成图无标识“裸奔”，使“有图有真相”的传统认知面临挑战，公众难辨真伪。

AI图片“随心定制”

记者在调查中发现，如今网络上存在多款AI制图软件，一些社交平台甚至自带AI生成图功能。只需上传相关人物、背景素材，输入一句指令，短短几秒便能生成一张几可乱真的AI图片，还可随意按照个人喜好对图片的细节和内容进行调整。

受访专家指出，如今AI生成图片或视频已很难从面部表情、肢体结构等方面识别真伪，如果不对内容作显著标识，很容易让人信以为真。

为调查平台与创作者对AI生成内容的实际标注情况，记者在一周时间内于6个主流短视频平台及社交平台的推荐页面，分别随机浏览并记录了100条AI生成或高度疑似AI生成的视频内容，并就其标注情况展开记录。

经统计，在随机浏览的视频中，约有60%的视频附有作者主动标注的AI创作声明；约30%虽未获作者标注，但被平台在显著位置标识为“内容可能为AI生成”。然而，在各大平台中，仍有部分高度疑似AI生成的视频，既无作者标识，也无平台提示。

以某短视频平台一个拥有30多万名粉丝的账号为例，其发布的内容多以萌宠、儿童和农村生活为主题。其置顶的3条视频，内容均呈现明显的AI生成痕迹：小狗轻易撞破机场玻璃；行李箱的颜色瞬间变化；数十名长相雷同、头身比例严重失调的婴儿一边走一边喊“奶奶”。然而，其中仅1条在视频右下角有AI创作标识。

在该账号的首页介绍中，明确附有其私人社交账号，并注明“可学习教学”。记者添加该账号私聊咨询后，对方承认所有内容均为AI生成，并表示“120元即可获取提示词以及全套教程，包教包会”“制作一条AI视频仅需十几分钟”。

靠AI生成内容引流卖货

此外，AI生成内容也被用于引流与商业推广。

“在××（地名），女儿已经过上了普通人里最好的生活”。近日，在某社交平台中，记者频繁刷到以此为标题的帖文。不同的帖文中，这个地名被换成北京、长沙、珠海等不同城市。

在这些帖文中，孩子父母的“人设”高度一致：爸爸做生意、妈妈在国企工作，父母学历均为985本/硕。虽然是不同账号发布的内容，但文案几乎完全雷同，均为教育经验分享或情感激励类内容，极易引发家长群体的共鸣。

这类帖文的配图普遍采用父母和孩子在校门前或景点前的合照。尽管账号的IP地址显示在不同地区，且背景不断变化，但画面中父母和孩子的形象高度一致。

同一张脸、同一个家庭，却有多段“精彩人生”？记者下载相关帖文照片后发送给多个AI软件询问“是否为真实图片”，得出的结论是“背景元素存在逻辑冲突、人物细节有AI特征，高度疑似AI生成图”。

然而，平台并未对这些图片添加提示AI生成的相关内容，图片本身也未包含任何水印标注。

记者进一步观察发现，这些账号用AI生成图塑造“成功家庭”形象的目的实为兜售课程。在某账号发布的帖文中，其表示“成功家庭”的核心秘诀是孩子的教育，进而围绕塑造孩子数学思维，展开长篇论述。评论区内不少网友留言“求推数学思维课”。记者咨询后，对方发来某数学课程商家账号，该账号经营的店铺内售卖标价为5450元的新数理课程。

值得注意的是，部分账号的主页直接关联某些线上教育机构的官方账号，并在后缀标注“产品体验官”“规划师”等身份标识。这类绑定关系显示出其内容与商业推广间的明确关联，客观上构成软性广告，但大多数家长往往难以辨别其营销性质。

除此之外，AI生成技术也被广泛应用于直播带货与商品展示。

某社交平台部分穿搭博主账号分享的内容中，AI生成特征较为明显：背景杂志上的人物和标题多为乱码、电梯场景中楼层按钮明显失真，并非真人实穿。然而，平台和博主均无任何关于AI生成的标注，博主还会通过评论区与用户积极互动：当粉丝询问服装链接时，博主会直接引导至商品橱窗；对于涉及身高、体重等个人特征的问题，博主也未提示任何AI生成属性。

在某购物平台上，多家销售睡莲的店铺均使用相同的商品宣传图，图片呈现明显的AI制作特征：背景失真、花朵呈半透明状。相关商品的买家评论最多达7万余条，而其评论区却几乎未见成功开花的反馈。在偶有的一些开花图片中，花朵数量和形态也与宣传图相差甚远。而上述商品宣传图中，均未出现AI生成图片的相应提示。

去标识就能“骗”过平台

为验证平台对AI内容的识别能力，记者购买了一款宣称“可去除品牌水印”的某AI软件会员，使用该软件可生成视频并进行去水印处理。随后，记者将有水印与无水印版本同时发布于4个主流社交平台，且均未主动标注AI创作。

调查结果显示，上述平台对有水印的版本均作出“含AI生成内容”的标识，对于无水印内容，仅有一个平台在后期补充了AI内容标识，并在后台私信发来风险提示外，其余平台均无提示。

为验证AI生成内容标识系统的实际有效性，记者对当前市面6款使用人数较多的生图软件展开调查。结果显示，虽然在内容生成时全部默认标注水印，但有些软件只要充值成为“高级会员”，即可解锁“去水印”功能。

此外，市面上存在多家提供低价去水印服务的商家。其中一款售价6元、宣传“终身包更新”的软件包，明确提示可去除图片与视频中的各类水印，并且不限使用时间和次数。

记者购买了该安装包进行实测，在按照要求完成软件安装后，发现仅需对水印位置进行简单涂抹，即可完整去除AI标识，且不破坏画面整体效果。随后，记者将去水印后的图片上传至4个社交平台进行验证，结果显示，这些原本带有AI水印的内容，在去除水印后均未触发任何平台的AI生成提示机制。

受访专家认为，平台在AI内容标识的落地执行方面还有很大的提升空间。不少网友把使用AI生成图视为娱乐或恶作剧，却忽视了背后可能存在扰乱社会治安、引发恐慌的风险。因此，在完善技术监管机制的同时，也要加强公众网络素养教育，提升公众对AI生成内容的辨识能力和社会责任意识。

漫画/高岳

AI生成内容真假难辨对平台治理提出挑战 专家建议

明确平台法律责任 提升审核识别精度

□ 本报记者 赵丽
□ 本报见习记者 丁一

近年来，生成式人工智能（AIGC）技术迅速发展。昔日需要整个团队协作完成的视频制作，如今仅凭一张图片、一句文字指令即可实现。然而，伴随技术普及而来的是滥用风险的加剧。纷杂的AI生成内容常常令人真假难辨，不仅干扰公众认知，对行业监管和平台治理也提出了新的挑战。

今年9月1日起，《人工智能生成合成内容标识办法》（以下简称《办法》）施行，意味着所有利用人工智能技术生成、合成的文本、图片、音频、视频、虚拟场景等信息，必须依法添加相应的声明标识。

然而在实际应用中，AI生成内容标注仍存在诸多难点。例如，用AI工具生成的视频，若用户有心将水印去除，平台往往难以有效识别其AI属性，导致大量未标识内容持续传播。

对此，《法治日报》记者采访西南政法大学公法研究中心执行主任杨尚东、中国农业大学法律系副教授薛铁成、中央财经大学法学院教授王叶刚以及北京航空航天大学法学院副教授王天凡，共同探讨应对之策。

记者：平台是内容传播和触达公众的关键环节，因此承担着监督和管理的重要责任。据观察，一些平台通常会在相关帖文上标注类似“疑似AI创作，请谨慎甄别”等提示。此类提示，是否足以认定平台已经尽到合理的注意义务？是否能使其自动免除法律责任？

王叶刚：不能据此当然认定平台已经尽到合理的注意义务。平台在性质上属于网络服务提供者，我国民法典侵权责任编对网络服务提供者的侵权责任具体规定了通知规则（避风港规则）、知道规则（红旗规则）。判断平台是否尽到了合理的注意义务，是否需要承担侵权责任，需要依据侵权责任编的规定进行判断，而不能以标注“疑似AI创作，请谨慎甄别”等提示语为由认定其尽到了注意义务。

王天凡：判断平台是否履行“合理注意义务”，需综合多方因素考量。比如，标识的显著性与有效性方面，提示语是否足够清晰、醒目，不易被忽略或裁剪；数字水印是否能抵抗压缩、剪辑等。防控措施方面，平台是否建立了与其技术能力和规模相匹配的、主动的侵权内容识别与过滤机制，以及在接到侵权通知后，平台是否能够迅速响应并采取删除、屏蔽等必要措施。另外，对于显而易见的、利用名人肖像进行虚假广告的深度伪造内容，法律对平台的“应知”要求会更严格。

记者：调查显示，很多AI生成内容引发的整蛊、造谣乱象，均是内容发布者有意为之，甚至故意绕开平台监管。对此，应如何加强管理？

薛铁成：必须明确，蓄意发布AI生成图片、视频整蛊甚至造谣的行为，可能涉嫌违法。若内容发布者明知其行为可能引发恐慌、报警、扰乱社会秩序，仍主动发布AI生成的图片、视频等误导性内容，其主观上可能构成谎报警情的间接故意，最终行为或被认定为报假警。根据治安管理处罚法相关规定，当事人将被处以行政处罚。若情节严重，还可能触犯刑法中的“编造、故意传播虚假信息罪”，须依法承担刑事责任。

加强对内容发布者的管理，要进一步压实平台责任。《办法》采用的是漏洞填补和责任加重的AI标识管理规则。在关涉AI生成内容发布的各个环节中，后一环节对前一环节负有检验和审核的责任。例如，提供网络信息内容传播服务平台发布AI生成内容时，既需要检验用户提供的AI生成内容是否有平台的隐式标识，还需要检验用户是否声明AI生成内容。如果均没有相关AI内容标识和声明，本平台检测到显式标识或者其他生成合成痕迹的，识别为疑似生成合成内容，也应当采取适当方式在发布内容周边添加显著的提示标识，提醒公众该内容疑似AI生成内容。

杨尚东：首先，平台应明确AI生成内容的管理规范。平台需制定清晰的AI生成内容管理规则，一方面强制要求用户对AI生成内容进行标注或主动声明；另一方面完善平台细则，针对恶意传播AI违规内容的视频及账号，依法依规采取提示、下架、封号等阶梯式处置措施。

其次，平台应加强AI生成内容的审核与标识。内容发布前，平台应升级审核机制，通过“技术检测+人工复核”的双重模式提升AI内容识别精度，精准筛选出可能扰乱社会秩序的内容及账号，并降低此类整蛊内容的传播权重。内容发布后，平台需在聊天、评论等交互环节，进一步强化AI生成内容的明确标识与风险提示，引导用户理性判断。

最后，平台应建立快速响应与事实核查机制。平台可搭建公众举报通道，及时发现审核疏漏的问题内容。针对可能扰乱公共秩序的信息，第一时间启动事实核查与应急响应，快速采取限流、删除等措施，防止内容扩散造成更大负面影响。

记者：针对AI生成内容标识的规范与治理，当前实践中还存在哪些核心问题，又该从哪些方面推进完善以保障落地效果？

薛铁成：当前相关法律制度仍缺乏AI服务提供平台和网络信息传播平台未履行AI标识义务的责任规定，难以建立对相关平台未履行AI标识义务的惩戒机制。在相关规定已经明确AI服务提供平台和网络信息传播平台在内容生成和发布环节义务的基础上，还应进一步明确其未履行义务的责任种类和承担责任的方式，以督促相关责任主体履行监管义务。

杨尚东：其一，健全AI生成内容管理的法规体系。未来，应在《办法》基础上进一步细化平台治理规范，明确政府监管职责、规范用户使用行为，强化法律执行效力，为生成、制作与传播合成内容的相关主体设定清晰的法律红线。

其二，构建跨部门协同监管机制。推动网信、公安等相关部门形成全链条监管合力，加强对生成式人工智能大模型的审核管理及其应用场景的监督，严格落实内容标识要求。同时，建立跨领域信息共享平台，提升监管联动能力与精准性，实现高效协同治理。

其三，强化网络空间法治宣传教育。借助各类媒体平台，向公众普及人工智能基础知识与AI生成内容的识别技巧，提升全民数字素养与技术认知水平。