2025.8.14 星期四
责编 薛亮 | 美编 小凯 | 校对 赖敏
江南晚报

有争议谁举证
如何自证“我不是AI”？

“手敲了一下午的剧本被认为是AI生成”“我引用的作者原话竟然被判定为AI生成”“只要说话像人机对话，都会被判定为是AI作答”……随着AI的广泛应用，一些“真人创作者”的作品也被平台误判为AI生成，甚至因此被限流、下架。

9月1日起，国家互联网信息办公室、工业和信息化部、公安部、国家广播电视总局联合发布的《人工智能生成合成内容标识办法》将施行。该办法明确，用户发布AI生成合成内容时应主动声明并进行标识，网络平台也有责任对AI生成内容进行标识，提醒用户注意识别。此举旨在防范利用AI实施诈骗、侵权等各类不法行为。那么对于被“误伤”的作品又该如何判定？

（新华社发 朱慧卿 作）

现象 真人创作却被误判为AI

画师“月饼”在某社交平台上传了1张原创绘画作品，谁知发布3小时后突然收到网友询问：“为什么这幅画显示是AI画的？”

“月饼”这才发现，自己这幅画作被平台标注为“疑似包含AI创作信息，请谨慎甄别”，但这条标注只有其他人能看到，“月饼”作为作者完全看不到。

“月饼”赶紧找客服申诉：“这幅画是我亲手绘制的，没有使用任何AI技术。”见AI标识一直还在，“月饼”无奈开始通过人工渠道申诉，并将自己的绘图过程、图层以视频方式整理出来。时隔几个月后，AI标识消失了。“由于作者看不到AI标识，我也没注意标识是什么时候撤的。”“月饼”告诉记者，“这幅画我画了30多个小时，从绘画软件中可以看到，这幅画含60多个图层。”

“原创被误判为AI，是对作者努力的否定。”“月饼”希望平台能建立明确、透明的AI标识撤销渠道，并向网络用户公开相应判断标准。

“月饼”的遭遇并不是个例。不少网络用户都表示，自己的画作、文案、视频也被标注过“疑似AI合成”，包括手绘的线稿、自拍照、风景照等。

“一些真正用AI生成的内容不一定会被识破，反倒真人原创的内容被当成AI。”有网友无奈地表示。除了打上AI创作的标识外，有些平台还会对作品进行隐藏、限流，甚至会对用户实行禁言等处罚措施。更让人困扰的是，不少平台目前缺乏专门针对“误判”的申诉机制。

探因 AI识别难度越来越大

今年初，一家大型互联网企业负责人在公开场合披露，在AI低质内容治理方面，目前普遍采用的策略是“用AI对抗AI”。约九成以上的审核工作由机器完成，但机器审核仍需人工不断校准，并持续为系统输入更准确的样本进行训练。

目前来看，AI检测并无标准，其准确率还有待提升。AIGCLINK发起人、零氪云联合创始人占冰强在接受记者采访时表示，目前各平台采用的AI识别技术不同，常见的是特征识别。“但因为AI生成内容与人类创作日益接近，且现在的人类输出内容中本身也包含了大量AI生成的原始资料，难以区分，导致该技术效果有限。”

而对文字内容的检测难度又高于图片、视频。“AI生成的图片、视频中不拟人、不拟合世界物理规律的地方比较多，鉴别相对简单一些。”一款AI检测工具的开发者告诉记者。

在多位受访者看来，随着AI模型的逐步发展，未来检测内容是否由AI生成的难度会越来越大。按照9月1日即将实施的《人工智能生成合成内容标识办法》，上游AI模型需要对生成内容、元数据文件添加标识；平台在检测内容时，应首先核验文件元数据中是否含有隐式标识，再看用户是否主动声明内容为生成合成内容，最后检测显式标识或其他生成合成痕迹，并根据不同情形打上相应标识。

国内某视频平台相关负责人表示，该平台为创作者提供了主动声明功能，鼓励创作者在发布AI生成内容时主动进行标识，对于疑似AI生成的内容，平台识别后会主动进行标识，提醒用户注意识别。该负责人表示：“相关技术需要不断迭代，才能提高识别的准确度。”在AI大模型能力逐渐提升的当下，占冰强认为，无需过度关注内容“是不是AI生成”，而是应该关注内容的独创性；也无需将人和AI作对比，而是应该对比“两个人使用AI生产的内容”之间的差异。

案例 一句评论引发首例AI误判案

北京互联网法院近期审结了全国首例平台运用算法工具进行AI生成合成内容检测识别引发的案件。

在某网络平台一条有关“高考后应该是学车还是打工”的提问下，网友李先生评论道：“打工并不能让你真的赚到多少钱，但可以开启你的新视角……如果你对学车感兴趣并且以后打算开车的话，可以在你最清闲的一个假期完成它……工作之后就没有太多完整的时间学驾照了。”

然而，这条两三百字的内容被平台判定因为“包含AI生成内容但未标识”而违规，这条内容被隐藏的同时，李先生的账号也被禁言一天。

明明是自己一字一句敲出来的内容，为何会被判定为AI生成？向平台申诉未果后，李先生选择起诉平台，要求法院判令平台撤销隐藏的内容和账号禁言一天的违规处理，并在后台系统中删除违规处理记录。

平台方则坚持认为，李先生发布的内容经机器识别为“包含人工智能生成”，人工复核也认为相关内容缺乏明显的人类情感特征。此外，平台还提供了一份公示的算法机制机理内容，作为算法决策判定的证据。

北京互联网法院审理此案后明确两点：第一，平台有权对涉案内容是否属于AI生成合成内容进行审查和处理；第二，平台的审查及处理结果应该有合理依据。

虽然被告平台提供了上述算法备案信息，但法院确认，该备案信息内容与本案争议没有关联性，同时，该平台没有对算法决策依据和结果进行适度的解释和说明。法院最终认为，在没有事实依据的情况下对用户进行了处理，应承担违约责任，因此判决平台删除对李先生评论的违规处理记录，撤销隐藏内容。

当平台判定用户发布的内容是AI生成，而用户却坚称是人类创作时，举证责任在谁？北京互联网法院综合审判一庭副庭长袁建华表示，在遇到此类问题时，一般情况下，用户具有初步的举证责任，来证明自己并非使用AI创作。

“比如，用户写作了大段文字，肯定存在一稿、二稿甚至多次修改的原稿；用户拍摄了一段视频，也肯定会留下源文件。”袁建华举例说。

上述案件的特殊性在于，李先生难以对两三百字的即时创作过程进行举证。“无法强求用户架起一个摄像机，来记录打字输入的过程。在此情况下，要求用户承担举证责任不具有现实可行性和合理性，平台就需要承担举证责任。因为平台主张的是对自己有利的事实，算法决策过程和判断结果都是由平台掌握的，平台需要对自己用算法判定出的结果的正确性和合理性进行举证，这样更符合公平原则。”袁建华说。

简而言之，在遇到类似情形时，原则上应该由用户承担初步的举证责任。提供初步证据后，平台应举证证明其使用算法工具进行判定的正确性或进行必要限度的解释说明。

袁建华进一步解释道：“在司法审查阶段，平台不能完全以商业秘密为由拒绝对算法的适度解释说明。算法毕竟是一个自动化决策过程，很容易形成‘算法黑箱’，算法失灵或误判可能会造成对公众权利的侵害。”

建议 完善复核机制 引入专家评估

在袁建华看来，AI生成合成内容标识作为对人工智能生成内容监管的有益工具和必要手段，会对净化网络内容环境和网络空间治理起到一定积极作用。但在算法识别AI的准确率没有达到100%的情况下，可能会存在平台误判的情况，需要平台前期进行处理，化解此类纠纷。

“在用户拿出相关原始证据的情况下，平台经过复核认为确属误判，需尽快去除对相关内容的AI标识、解除对相关用户的处罚措施。”袁建华建议，平台应完善复核机制，建立便捷、有效的申诉渠道。如有条件，还可在算法识别基础上引入专家评估机制，综合判断内容是否为AI生成。

此外，网络用户在使用AI工具进行创作并将相关内容发布到网络平台时，应依照诚信原则如实进行标识，以此维护健康良好的互联网环境。网络用户在发现内容被误判为AI创作后，应积极与平台沟通，并尽量提供完整的创作证据。

（北京晚报）

画师“月饼”的作品被打上疑似AI制作标识。