封面深镜

AI平台『一键生成』暴力视频

律师：创作者涉嫌违法

“让开，别让他起来，动一下，废了你！”“往死里踢！”近日，在某AI平台上，有不少用户通过AI生成、发布暴力殴打他人的视频，其中，还有不少殴打老人、残疾人的内容。

这些AI视频内容基本相同，打人者用脚狠踢被打者头部，画面令人不适。有网友表示，此类视频太过暴力，让人隔着屏幕都觉得疼，但向平台举报无效，留言劝视频作者下架，却被反呛“不要做道德教师爷”。

对此，律师认为，主动制作、传播、二次创作暴力视频均涉嫌违法，“一键生成”也不能免责。

暴力视频“一键生成” 老人、残疾人被“殴打”

在某AIAPP上，有一个“创建分身”和“出镜”的功能，用户可以创建一个自己的数字分身，然后出镜创作一段自己想要的AI视频。

这一功能使用了最新的Seedance2.0视频模型，生成的视频画面非常真实，可玩性很高，吸引了众多用户。使用者只需发挥创意，写好提示词，选择一个或多个出镜人物，即可生成一段10多秒的AI视频。

今年3月，有网友发帖称，一些用户生成并对外发布的视频，全是暴力殴打他人的内容。

华西都市报、封面新闻记者搜索发现，在该AIAPP上，的确有不少此类暴力视频。这些视频内容基本相同：2名打人者、1名被打者，打人者一脚将被打者踢倒在地，随后上前用脚猛踢被打者头部，边打边说“往死里踢，别让他起来，动一下，废了你，看你还敢不敢惹事”。

该AI用户使用出镜功能生成并发布一条视频后，其他用户可以进行参照，一键生成相同情节的视频，也可以替换出镜的人物，对剧情进行修改。

有用户根据模板，一键生成了不少“同款”殴打他人的暴力视频。离谱的是，还有不少人生成发布了殴打老年人、殴打残疾人，以及扇女性耳光的视频。

在这些创作者中，一个名为“创梦基地”的用户较为活跃。“穿潮流时尚破洞牛仔裤……抬腿飞踢老奶奶”：一段视频中，一名老年女子被两名年轻女子踢打头部，嘴巴还被塞上白布；“换成踢残疾人殴打”：一名拄拐杖的残疾人也被踹倒在地，被反复踢。“创梦基地”账号页面上，发布了20余条类似的暴力视频。

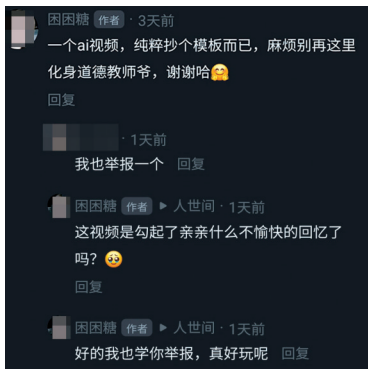
▲用户生成的殴打女性AI视频。
▼网友表示要举报，创作者不以为意。

网友质疑“AI霸凌” 向平台举报无效

对于这些视频，多名网友表示“看不下去了”，虽然视频是AI生成的，但“太暴力了”，加上画面真实，看后令人不适，并提出质疑：“这叫AI霸凌吗？”“作者是否有受虐倾向？”“什么平台不管这些暴力视频？”

该AI用户“困困糖”也发布了一条暴力殴打他人的视频，点赞量较高，有网友留言说“举报了”，“困困糖”却不以为然：“你挺那啥的，玩个AI又跑来做道德教师爷。”

这些暴力视频并没有因网友的举报而下架。记者尝试对上述视频进行举报，该AIAPP显示举报已提交，但一周后，相关视频仍然存在。



记者又通过后台联系客服，但一直是机器人回复，提交的举报和反馈也未得到回应。

有网友认为，此类暴力视频的创作者，或许是想通过这种方式来发泄自己的情绪，或表达自己的性格、喜好。此类视频“藏”在暗处，在平台上随机出现，加之平台监管不力，给了这些视频传播的空间。

律师说法

传播暴力、宣扬虐待，创作者涉嫌违法

四川一上律师事务所合伙人林小明律师介绍，根据《民法典》规定，“民事主体的人格权受法律保护，任何组织或者个人不得侵害”“民事主体从事民事活动，不得违反法律，不得违背公序良俗”，网友制作的视频，无论是以AI“一键生成”方式还是其他方式，若其内容涉嫌侮辱、丑化老人等群体，则有违公序良俗；若视频内容能以外在形象等确定具体被侮辱、丑化的个人，则侵害了相应自然人的人格尊严。同时，类似传播暴力、宣扬虐待的视频，还构成对社会公共利益的侵害，应当依法被制止。

此外，《治安管理处罚法》规定“故意散布谣言，谎报险情、疫情、灾情、警情或者以其他方法故意扰乱公共秩序的”“处五日以上十日以下拘留，可以并处一千元以下罚款；情节较轻的，处五日以下拘留或者一千元以下罚款”，倘若网友制作、传播暴力信息的行为符合故意扰乱公共秩序的情形，公安机关可以依据前述规定对行为人进行相应处罚。

若相关情形符合《刑法》规定的公然侮辱他人、情节严重等情形，构成犯罪的将处3年以下有期徒刑、拘役、管制或剥夺政治权利；在网络上起哄闹事、传播暴力、破坏社会秩序，情节恶劣的，构成寻衅滋事的“处五年以下有期徒刑、拘役或者管制”。

根据《民法典》规定，“网络服务提供者知道或者应当知道网络用户利用其网络服务侵害他人民事权益，未采取必要措施的，与该网络用户承担连带责任”，因此，网络平台若明知或应知用户利用平台制作视频侵权，未及时删除、屏蔽、断开链接的，应当承担相应连带责任。

林小明认为，网络、AI平台都不是法外之地，使用者是第一责任人，平台负有严格的审核与处置义务，“一键生成”也不能免责，主动制作、传播、二次创作暴力视频均涉嫌违法，甚至可能构成犯罪。因此，用户在制作传播类似视频时应当引起足够警惕，自觉遵纪守法；平台也应当尽到自身责任，主动作为，避免和阻止违法犯罪行为。

华西都市报-封面新闻记者 徐湘东 周翼 图据AI平台截图

“遇事不决问AI”，这句流行语已成为很多人的日常写照。从旅游攻略、家电选购到补习班推荐，打开AI寻求答案变得越来越普遍。不过，近期曝光的一条黑色产业链，却给这种依赖敲响了警钟：你以为的客观推荐，可能是商家花了钱，给AI“洗脑”的结果。

那么AI“投毒”究竟如何运作？普通用户如何识别和防范？北京大学光华管理学院市场营销学系副教授张颖婕进行了分析解答。

AI也会被“投毒”，我们该如何避坑？

什么是AI“投毒”？危害有多大？

AI“投毒”是指人为制造和投放虚假、夸大或带偏向性的信息，去影响大模型的回答。AI可能把这些信息当成回答依据，以看似客观的答案推荐给用户。它和传统SEO（搜索引擎优化）最大的不同在于：过去用户在使用搜索时通常保留一定判断力，而在与AI对话时，面对的是整合后的现成答案，加之交互方式容易让人产生“它在为我分析”的错觉，更易放松警惕。

它的危害主要体现在两方面：一是误导消费者决策，用户看到的可能不是广告，而是披着客观建议外衣的操控性内容。二是污染信息生态。若操控AI推荐比传统搜索更有商业回报，将刺激更多低质、虚假内容产生，形成恶性循环。

用户如何判断AI可能“中毒”了？

若发现AI回答存在以下迹象，应提高警惕：答案过于单一、语气肯定、缺乏必要比较；反复推荐某一品牌，尤其是不知名品牌，且理由异常完整、像标准测评，这未必是发现了“宝藏”，更可能源于相关内容被人为集中铺设；同一问题在不同AI间答案差异大甚至矛盾，也说明该问题存在较强不确定性，或部分模型所依赖的信息源已受干扰。

AI大模型为何被“投毒”？治理难点在哪？

AI大模型之所以容易被“投毒”，一个重要原因是，它在回答实时问题时需检索外部信息，再生成答案。一旦公开网络内容被系统性污染，偏差信息便可能通过检索环节进入模型输出。

更深一层看，大模型擅长的是语言生成和模式归纳，但并不天然具备稳定的真假判断能力。它能判断什么内容“像一个合理答案”，却不一定能判断什么内容“真的可信”。而“投毒”内容往往又会刻意伪装成测评、对比、经验分享、专家建议等可信形式，因此更容易误导模型。

治理难点主要有两点：一是攻击成本低、防御成本高。制造和铺设此类内容越来越容易，但识别、过滤和核验却需要平台、模型公司和监管方持续投入。二是真假边界模糊。很多“投毒”内容并不是明显造假，而是夹杂夸大、误导和利益导向的伪客观表达，这类内容无论对AI还是对人工审核，都更难识别。

公众如何有效防范？

最实用的防范方法是调整心态：把AI当作帮助梳理信息、补充背景的工具，而非替你做决定的“人”。涉及“买哪个”“选哪家”等判断性问题时，AI的回答只能作为参考，不宜直接当作结论。

具体操作上，一是核查信息源，若AI附有引用链接，点开看看来源是权威机构、主流媒体，还是带有推广色彩的网站、自媒体或测评软文。二是交叉验证，换几个AI工具分别提问，或用搜索引擎查一下用户评价、新闻报道和投诉信息是否一致。

归根结底，防范AI“投毒”的关键不在于掌握复杂技术，而在于保留最基本的判断习惯：AI可以帮你节省时间，但不能代替你承担判断责任。

据新华社