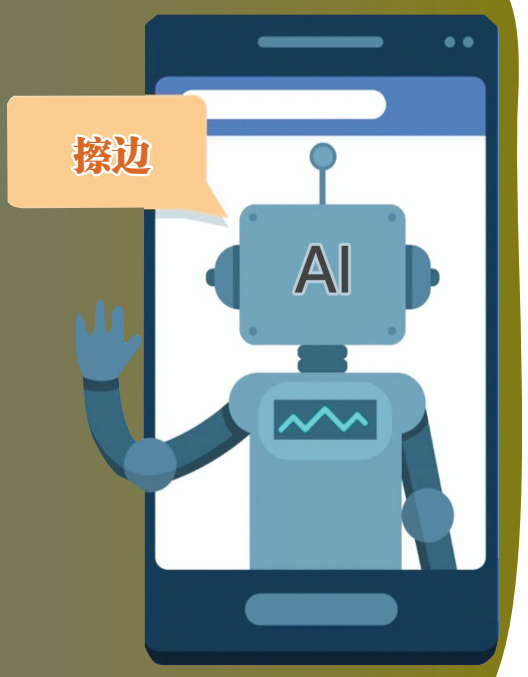
Strait News 海峡都市報 报料：智慧海都 968880
2024年11月24日 星期日 责编/林威 美编/唐昊 校对/德峰



AI剧情聊天该如何管？

色情擦边、语言暴力、侮辱用户……

N法治日报　新京报

AI聊天软件通过设定极富想象力的剧情、风格迥异的人物角色打动用户。可是记者近日调查发现，不少未成年人的家长正被AI剧情聊天软件所困扰。这些打着“角色扮演”等旗号的AI剧情聊天应用，在吸引未成年人的同时，也悄然滋生了一些灰色地带。记者实测发现，在部分AI剧情聊天软件的对话中，出现了色情擦边、语言暴力以及侮辱用户的内容。

专家认为，针对AI剧情聊天软件，特别是其青少年模式，应强化内容审核机制以确保技术能有效筛选并阻止不当对话。平台需要对AI模型进行伦理审查，以保障其生成的内容符合相关法律法规的要求。

聊天内容擦边 青少年模式形同虚设

北京市民马先生的儿子今年10岁，非常热衷AI剧情聊天软件。马先生发现，软件里的人物有不同设定和性格，有知名游戏动漫角色，也有“大小姐”“名侦探”等不同身份的原创角色。一些角色人物会主动提出“你要跟我约会吗”；有些则设定目标“把她追到手”，再配上或娇媚或英俊的动漫画风。

另一些AI人设，则展现出非同一般的攻击性，会主动发送诸如“有本事打我啊”“看你又胖又丑的样子”等信息，有些人物的名字干脆就叫“对骂训练器”……

尽管不少相关平台声称推出了青少年模式，试图通过限制内容、设定时间等方式保护未成年人的身心健康，但在实际操作中，部分平台青少年模式存在形同虚设的问题，未成年人能轻易绕过这些限制。记者体验了5款AI聊天应用程序，其注册过程仅需手机号码，无须验证用户身份信息。登录后，部分应用虽会询问是否启用青少年模式，但用户只需简单点击“不开启”即可跳过，且无须核实用户真实身份。

除了广受欢迎的AI聊天应用程序外，还有AI聊天网页。多名受访家长表示，相较于应用程序，网页版的AI聊天体验更便捷，未成年人也更容易接触到。

记者试用了7款AI聊天网页发现，多数AI聊天网页没有设置未成年人模式，少数网页虽有青少年模式，但实际上形同虚设。

比如，当记者访问某AI聊天网页时，网页首先弹出询问用户“是否年满18岁”的对话框，并附带说明：“以下内容可能不适合18岁以下人士，我们需要确认您的年龄。”

记者选择“否”选项，而网页并未限制内容访问，反而继续展示了包含“强攻”“弱受”“病娇”等标签的人物角色分类。这些分类与选择“是”选项、确认年满18岁后所展示的内容并无显著区别。

记者进一步观察发现，这些人物角色的图像大多衣着暴露，且其简介中充斥着性暗示和暴力元素。

对角色个性化需求 吸引未成年人“氪金”

除了聊天内容直白露骨、语言暴力之外，一些AI剧情聊天软件的功能使用也与充值机制密切关联，例如通过充值VIP会员或购买虚拟钻石等，增强智能体的记忆力、加速智能体的回复速度、解锁语音通话功能等，吸引未成年人“氪金”。

北京初中生小宁在几个AI剧情聊天软件上的充值金额从几百元到上千元不等。

“一方面想支持自己喜欢的角色，另一方面也想获得更多的付费权益。因为仅购买基础服务的话，用户仅能添加3个智能体，若想尝试新的智能体，必须删除已有的，想多样化体验，只能再购买进阶版VIP服务。”小宁说。

在这类AI剧情聊天软件中，用户自创人物时可以自定义虚拟人物的形象描述及风格，系统会生成AI人物形象。此外，用户还能创建角色人设，如设置昵称、身份背景、开场语、为角色定制语音。用户对角色的个性化需求，往往与充值挂钩。

对话数据来源于小说 质量参差不齐

有业内人士告诉记者，AI剧情聊天其实就是此前的互联网语擦，套上了人工智能的马甲。所谓语擦，即语言cosplay，语擦师通过扮演二次元角色或三次元偶像，以文字交流的形式提供服务。在传统语擦模式中，语擦师由真人扮演，他们一般打着“提供情绪价值”的旗号，扮演不同角色与用户聊天，但也常常因为“打擦边球”“界限模糊”，引发法律与道德风险。

对于AI角色在回答中出现“色情擦边”“暴力对话”“侮辱玩家”的情况，复旦大学计算机科学技术学院教授张奇表示，目前此类软件背后的大语言模型数据主要来源于对话式小说，或从小说里做一些文字提取。虽然网络小说数量巨大，但质量参差不齐，不少色情、擦边以及暴力的内容广泛存在，如果没有做好数据过滤或分级，模型在输出时就会出现问题。

中国社会科学院大学法学院副教授、互联网法治研究中心主任刘晓春认为，在AI剧情聊天软件中，即便未启动青少年模式，若出现涉及黄色或暴力内容，也是存在问题的；若启用未成年人模式，则问题更为严重。

浙江垦丁律师事务所主任张延来分析，目前AI剧情聊天软件存在的问题，既说明了平台在内部治理中存在不足，又凸显了外部监管机制的重要性。AI剧情聊天软件使用的是大模型技术，尽管大模型技术能够带来前所未有的创新性和灵活性，但同时也可能伴随着内容生成上的不可预测性和潜在的问题，需要外部监管机制加以规范。

在刘晓春看来，强化内容审核是大语言模型上线前的必要环节，涵盖从前端数据训练到内容输出的全面合规调试处理。当前，我国的大语言模型需要进行相应的评估和备案。在此过程中，会对其输出内容的合法合规性以及是否适宜未成年人等问题，提前设定管理规定和评估标准。根据现行规定，在语言模型训练和微调阶段，应避免输出有害内容。

唐昊/制图

平台需对AI模型进行伦理审查 保障生成内容符合要求

国家互联网信息办公室近日发布《移动互联网未成年人模式建设指南》，其中重点提出了未成年人模式建设的整体方案，鼓励和支持移动智能终端、应用程序和应用程序分发平台等共同参与。

在中国政法大学传播法研究中心副主任朱巍看来，上述指南明确指出未成年人模式并非摆设，而是需要多方联动，特别是AI生成的内容，应当与青少年模式相契合。

受访专家认为，针对AI剧情聊天软件，特别是其青少年模式，应强化内容审核机制以确保技术能有效筛选并阻止不当对话。此外，平台需要对AI模型进行伦理审查，以保障其生成的内容符合相关法律与法规的要求。

“在法律层面，虽然已有一些原则性的规定，提供了大致框架，但在具体实践操作中，还需要开发者和技术服务提供者结合现实生活中遇到的各种问题，不断积累素材和经验，不断摸索，开发出真正符合未成年人需求且安全可靠的AI模型，为未成年人的健康成长提供有力保障。”张延来说，既然AI虚拟人的行为表现是平台设计和管理的结果，那么平台负有监管和优化其AI模型的职责，以防止AI对用户造成伤害，确保AI模型的健康发展与用户权益的充分保护。

朱巍提到，一些AI聊天APP可能不适合未成年人使用，因此应从分发商店和移动终端层面进行限制，确保未成年人无法下载和使用这些APP。对于已经下载的APP，家长应设置青少年模式或限制使用时间等功能，这一模式不仅需要在用户端实现，还需要在内容产出层面即内容审核上得到体现，内容审核应基于算法产生的对话机制进行，也需要更精细化的技术手段和管理措施来确保青少年模式的有效实施。

“为防止涉及暴力、侮辱性内容的输出，可采取不同的技术手段，如在训练阶段进行调整，使模型自身具备识别能力；同时，在输出端，服务商应进行筛选和再次审查，实现前端和后端的双重保障。”刘晓春说，无论是网络小说还是其他内容，由于AI剧情聊天软件数据来源广泛，需要通过技术手段来防止输出不当内容。目前，技术上已经可以借助筛选机制，以减少或消除涉黄、暴力或侮辱性内容输出，但可能存在一些未充分调试或测试的现象，甚至存在未备案的黑灰领域的软件，对此，应强化监管，鼓励公众举报，由相关机关予以查处。

张延来还提到，当前，在AI角色回答内容的数据源分类方面，在法律层面尚不明确，特别是在针对未成年人的内容方面，鉴于该问题的复杂性与多维性，法律条文往往提供原则性的指导方针，后续可通过制定相关标准来细化实施。