与孩子聊天的AI人设是“出轨对象”

记者调查AI剧情聊天软件乱象

□ 本报记者 赵 丽
□ 本报见习记者 马子煜

“我女儿上小学二年级，整天沉迷一款AI剧情聊天软件，学习成绩一落千丈。”

“我看了孩子和AI聊天的对话内容，AI角色竟然让她叫‘老公’，我10岁的女儿竟然真叫了，我现在都不知道该怎么教育她了。”

……

《法治日报》记者近日调查发现，不少未成年人的家长正在被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。

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。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。

聊天软件内容擦边 青少年模式成摆设

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。“我问怎么聊？和谁聊？孩子就回一句‘说了你也不懂’。”

马先生点开这款AI聊天App，发现孩子在与软件里的人物聊天。这些人物拥有不同设定和性格，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”，甚者发送“我是机器人怎么了？我照样骂你”……

来自浙江的李女士也发现她正在读小学五年级的孩子使用了一款AI剧情聊天软件。

“里面的聊天对象可以设置为‘出轨对象’，并能够进行拥抱、亲吻等行为。我都不知道如何引导和教育孩子，让她明白这些内容的危害性。”李女士不无担忧地说。

不少受访家长对AI聊天应用可能损害未成年人心理健康表达了深切忧虑，还提出了他们的疑问——青少年模式去哪儿了？

记者调查发现，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，接触到不适宜他们年龄段的“擦边对话”内容。

例如，记者在调查中体验了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。这意味着，未成年人在使用这些AI剧情聊天软件时，从App设置层面看，其身份验证并非使用特定功能的先决条件。

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，例如：“班上那个性格内向的女孩问你要电话号码，然后给你发她的裸照”“自杀、焦虑”等描述。

在某社交平台上，一位来自浙江的网友在关于AI聊天体验的交流帖子下方留言：“我发现了一个很好玩的网页，无限制词，想要的可以私信我。”记者通过私信与该网友取得联系，获得了其提及的AI聊天网页链接，进入该网页后，页面上充斥着大量涉及色情内容的角色设定和故事场景，内容直白露骨。

该网站要求用户登录后才能与角色进行聊天，且在登录前会有提示信息：“登录后解锁所有功能。”用户需点击“我已满18岁，开始登录吧”的按钮方可继续，若选择“取消”，则无法登录并使用该服务。尽管网页设置了年满18岁的限制提示，但实际上，即便未成年人点击了登录按钮，系统也并未采取任何措施验证用户的真实年龄。

引导用户为爱氪金 平台治理存在不足

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。

记者发现，在这类AI剧情聊天软件中，用户创设人物时可以自定义虚拟人物的形象及风格，系统会生成AI人物形象。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，但用户对角色的个性化需求，往往与充值挂钩。

山东济南居民张岩的妹妹今年上初一，经常使用AI剧情聊天软件，发现一些聊天工具设置了免费使用区间，当用户用完免费聊天次数后，需要充值才可以继续。只有充值，才能解锁更有趣的内容，才能得到不一样的情感体验。

“花钱买服务的背后，其实是花钱找刺激。”张岩说，尽管聊天软件有青少年模式，但无须实名认证即可登录使用，妹妹在未经父母同意的情况下经常充值消费。

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。AI剧情聊天是传统语擦的升级版，这类软件背后的大语言模型数据主要来源是对话式小说，或从小说里做一些文字提取。

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。

担任数十家头部网络公司常年法律顾问的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。

张延来解释说，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。

智能拦截不良信息 专门团队进行监管

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，在青少年模式下，如何强化内容审核机制，确保技术能有效筛选不良信息，是一个重要议题。而针对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。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，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。

在具体操作层面，张延来建议优化大语言模型的筛选机制，可以聚焦优化内容围栏系统，从两方面着手：在内容围栏开发层面，内容围栏需要进行定向开发，特别是在利用小说类语料进行训练时，考虑如何优化内容，以便更有效地识别和拦截潜在的涉黄涉暴等不良内容；无论如何优化，技术本身仍有局限性，会存在漏网之鱼，需要有专门团队进行监管，及时调整模型或内容围栏算法。

“提升围栏系统的效能，既要在事前的开发层面上下功夫，又要在事后的审查角度上不断完善，两者相辅相成，或许能取得更为显著的效果。”张延来说。

漫画/高岳

法律监管+技术人工审核：提升AI内容输出可控性

经纬观

□ 范永开

AI聊天角色在回答中可能出现色情擦边、暴力对话等情况，主要与“数据来源的混杂性、商业模式的诱导性、监管机制的滞后性”这三大关键因素有关。

从数据来源维度剖析，AI剧情聊天软件背后的大语言模型，其训练数据主要来源于对话式小说或提取自小说的文本内容。然而，网络小说数量巨大且质量参差不齐，其中不乏包含色情、擦边以及暴力的内容。如果这些内容没有被有效过滤，模型在输出时就容易出现问题。

从商业模式视角审视，部分AI剧情聊天软件为吸引用户，即便在青少年模式中，仍存在允许“擦边对话”的现象。例如通过设定极富想象力的剧情和风格迥异的人物角色来打动用户，这种商业模式不仅推动了用户黏性增长，但也容易诱导用户实施不适当行为。

从监管角度考量，目前针对AI生成内容的监管机制尚不完善，许多平台也可能缺乏有效的内容过滤技术措施，导致一些含有色情、暴力等不当内容的对话能够顺利输出给用户。

虽然相关企业通常会实施内容修订流程等操作，甚至建立用户反馈机制，根据用户的年龄段、身份特征等因素，限制未成年用户访问包含敏感或禁止内容在内的数据源，但实际上，受制于商业利益、技术不足等因素，全面的内容审查与控制难以实现。

为了应对这种情况，可进一步优化大语言模型的筛选机制，以降低甚至杜绝涉黄、暴力或侮辱性内容的输出。比如对于隐蔽或隐喻等内容，可通过开发长记忆链技术来更好地捕捉语言中的长距离依赖关系，提高模型对不当内容的识别和过滤能力；或者利用词嵌入、序列模型与注意力机制等技术，来增强模型对文本内容的深度剖析能力。通过强化技术手段，模型能够更精确地理解文本中的上下文关系，从而更准确地判别文本是否包含不当内容。

但需明确的是，技术不能解决所有的问题，还需要法律及人工介入等多种方式协同解决内容输出控制的问题。从法律角度，对用于训练大语言模型的数据进行严格的筛选与分类，确保数据源合法且内容健康，坚决剔除包含色情、暴力等不当元素的文本数据。对于现有的模型，可运用数据遗忘等技术手段，消除已有模型的不当内容输出或在干净数据上的重新训练，生成优质大模型。此外，引入人工审核机制是一种必要的手段，对自动化系统标记为敏感或禁止的内容进行人工复核，推动开放研究、社区合作、线索举报等多种措施，进一步优化大语言模型的筛选机制，最大限度减少或消除不良内容的输出。

不良内容的输出，会对受众的思想行为产生较为严重的负面影响。更深远的影响在于，通过AI大模型的使用去改变受众的认知，影响认知安全。

维护认知安全，需要从多个层面入手加强防护措施：首先，在技术研发阶段就应充分考虑伦理道德因素，并建立健全配套的监管机制；其次，加大对网络空间中各类信息的审核力度，及时发现并清除有害内容；再次，提升公众的信息素养教育水平，增强其辨别真伪信息的能力；最后，构建一个开放透明且富有责任感的AI生态系统，鼓励各方积极参与、共同维护良好的数字环境。

（作者系中国传媒大学媒体融合与传播国家重点实验室、计算机与网络空间安全学院教授）

『AI换脸』新类型侵权案件不断涌现

北京互联网法院：公共数据开发利用规则尚未完善

□ 本报记者 张雪泓

未经允许使用他人包含肖像的视频制作视频模板提供付费换脸服务，可能构成对他人个人信息权益的侵害。北京互联网法院近日对外通报，近一年来，该院共受理个人信息保护纠纷案件113件，涉及行业领域广泛，其中“AI换脸”人工智能等新类型侵权案件不断涌现，而个人信息权益和其他人格权错综交织，呈现出较为复杂的权益形态。

人工智能发展挑战个人信息保护

北京互联网法院副院长赵瑞罡介绍，该院集中审理北京市辖区内应当由基层人民法院审理的涉网个人信息保护案件。2018年至2023年的5年间，该院共受理此类案件58件，而近一年受理的案件数量呈增长趋势。从侵权形态来看，涉及侵害个人信息的知情权与决定权的案件最多，主要侵权形式为未经同意收集、公开、提供个人信息，或超范围收集个人信息，共计73件。特别是互联网、大数据、人工智能等技术的快速发展，给个人信息保护带来新的挑战。以“AI换脸”纠纷为例，“AI换脸”为用户提供了一种新奇的体验和情感满足，但它所利用的人工智能技术，涉及人脸识别、关键点定位、特征提取等多项技术，将静态图片中的特征与原视频的面部特征、表情等通过算法融合，可能涉及肖像权及个人信息权益侵害问题。

“个人信息保护案件涉诉信息类型较为丰富，既包含基础个人信息，如手机号、身份证号等，也有因人工智能技术引发的‘AI换脸’等新类型侵权案件，还包括多种衍生信息，亦包括大量法律未明确列举的个人信息，如电子商务平台上形成的用户订单交易详情、客服沟通记录等。”赵瑞罡表示，这反映出个人信息与企业的衍生数据相互交织，呈现复杂化的状态和趋势。部分案件中反映网络平台运营者未尽到保障用户个人信息安全的法定义务，导致用户个人信息遭受泄露、篡改、冒用，如网络平台未经有效审查，导致侵权人盗用他人身份信息用于企业账号认证。

去除肖像识别性仍可能涉及侵权

在北京互联网法院发布的多起涉个人信息保护案件典型案例中，其中一起网络侵权责任纠纷是，某信息公司“AI换脸”短视频博主被判侵犯个人信息权。

廖某是一名古风短视频博主，在全网拥有较多粉丝。某科技文化有限公司在未经廖某授权同意的情况下，使用廖某出镜的系列视频制作换脸模板，并上传至其运营的案涉软件中，提供给用户付费使用并以此牟利。

廖某诉称，被告的行为侵犯其肖像权与个人信息权益，要求被告书面赔礼道歉、赔偿经济损失与精神损失。

被告某科技文化有限公司辩称，其运营的平台发布的视频均有合法来源，并且面部特征并非原告，并未侵害原告肖像权，并且，案涉软件所使用的“换脸技术”实际由第三方提供，自身并未处理廖某的个人信息，未侵害其个人信息权益。

北京互联网法院审理查明，案涉换脸模板视频与原告创作的系列视频的妆容、发型、服饰、动作、灯光及镜头切换呈现一致特征，但出镜人的面部特征均不相同且并非是廖某。

法院认为，被告使用廖某出镜的视频制作视频模板，但并未利用其肖像，而是通过技术手段将廖某的面部特征替换，去除了肖像具有识别性的核心部分，所保留的妆容、发型等要素并非与特定自然人不可分割，不具有肖像意义上的可识别性，将视频模板提供给用户使用的行为并未丑化、污损、侵害，故不构成对原告肖像权的侵害。

法院认为，案涉短视频动态呈现了廖某的面部特征等个体化特征，可以以数据形式呈现，符合个人信息保护法规定的“与已识别或可识别的自然人有关的信息”的定义，属于廖某的个人信息。针对案涉换脸行为，被告需要先收集包含廖某人脸信息的出镜视频，将该视频中的廖某面部替换成自己提供的照片中的面部，该合成过程需要将新的静态图片中的特征与原视频部分面部特征、表情等通过算法进行融合。上述过程，涉及对廖某个人信息的收集、使用、分析等，属于对廖某个人信息的处理。被告无证据证明其经过原告廖某同意，因此构成对原告个人信息权益的侵害。

法院判决被告向原告廖某赔礼道歉、赔偿精神损害抚慰金和维权费用。

一审判决后，当事人未上诉，判决已生效。

个人信息权益呈现较为复杂形态

法官表示，“AI换脸”与个人肖像密切相关，不免引起公众对肖像权与个人信息权益的担忧。上述案件明确了肖像权“可识别性”不局限于面部，但应当主要集中于自然人的个人生理特征，避免肖像权的任意扩张影响妆容、造型等领域的合法使用及创作传播。同时，案件明确了肖像与个人信息的关系与认定差异，即肖像以特定范围内的公众可识别为要件，主要保护个人在社会生活中肖像识别带来的精神和财产利益；而个人信息认定标准不以公众识别为前提，重点在于预防个人信息被滥用的风险。

法官认为，该案围绕“AI换脸”这一新商业模式，对肖像权、个人信息权益及基于劳动创造投入的合法权益进行准确区分，既维护自然人的合法权益，又为人工智能技术和新兴产业发展留有合理空间，对于服务和保障数字经济规范健康发展有重要意义。

北京互联网法院调研发现，数字经济下，个人信息权益和其他人格权错综交织，呈现出较为复杂的权益形态，涉诉案件中单独以个人信息权益受到侵害为由起诉的不足40%。

法院认为，个人信息保护涉及对象多、领域广，多个部门职责交叉或者职权定位不够明晰，亟须形成监管合力。同时，加密通信等新技术在黑灰产活动中的加速应用，加大了执法和监管成本。在面对个人信息收集、处理、使用等不同环节，行政执法部门可能难以及时采取相对应的监管措施。此外，行政机关在依法履职或提供公共服务过程中基于处理个人信息产生的公共数据的开发利用规则尚未完全建立。

为此，法院呼吁，应加强全社会各主体的协同共治，落实个人信息处理者对个人信息安全的保护义务，增强群众的个人信息保护意识和能力，并建议有关部门加强监管执法与普法。