擦边、文身……AI生成涉军不良形象

司法如何为人工智能划定“安全轨道”？

聚焦

□ 首席记者 季张颖

叼着烟刻着文身的“军人”，身穿擦边暴露的女式“军装”，身后还停着明显带有中国部队标识的装甲车……在手机应用商城内一款下载量达上百万的APP上，通过AI技术肆意拼接元素，丑化亵渎军队、军人形象，违背军服穿着规范的图片竟然可以随意生成，上海检察机关循线追踪，及时对这一人工智能领域新类型案件展开“问诊”。

生成式人工智能等前沿技术在催生产业新活力的同时，也带来了错综复杂的新挑战，从换脸变声、“一键脱衣”等AI滥用现象，到算法黑箱、数据语料安全，各类问题逐步暴露出来。

当前，上海正在加快推进建设人工智能高地，围绕司法服务助力行业高质量发展这一关键命题，近日，记者采访了法律及行业内专业人士，试图理清AI应用中的法律边界，探寻可行的治理路径。

语料库不设限 AI生成涉军违法不良信息

今年8月，徐汇区人民检察院在履行公益监督职责过程中发现，一家注册在本市的大模型企业存在未严格落实涉军信息内容管理主体责任的问题。

“当时，中央网信办针对AI技术滥用问题开展专项行动，我们在此背景下对辖区内多家备案的大模型企业展开初步调查。”案件承办人、徐汇检察院公益检察室检察官助理陈思琦告诉记者，就在排摸过程中，检察院发现在一家大模型企业开发的一款AI应用程序和服务网站中，只要输入相关提示词，就能生成丑化军人形象、违背军服穿着规范的图片，其中不乏一些色情暴露、明显有违军人形象的画面。

“这意味着，企业研发的这款大模型存在训练数据及生成内容安全隐患，管控策略存在缺失，防控体系存在漏洞。”陈思琦告诉记者，“只要用户输入相关提示词，AI软件和网站就能支持用户生成有违军容风纪的图片，这无形之中增加了风险。”

在查实涉案企业未严格落实涉军信息内容管理主体责任，可能侵害国家利益和社会公共利益后，徐汇检察院立即在当月以行政公益诉讼立案。“在此之后，我们与上海军事检察院依托军地检察机关协作机制开展联合办案，围绕履职难点，召开了专题联席会议，在界定损害事实、协助调取证据、准确适用法律等方面达成了一致。”徐汇检察院副检察长张爱菁介绍。

“人工智能行业正在发展阶段，在高速发展中必定也会带来一系列的挑战，我们检察履职的触角要跨前一步，从而规范人工智能服务和应用，促进行业健康有序发展。”张爱菁告诉记者，在这一“重治理”的办案思路指引下，检察机关向相关职能部门制发检察建议，督促其履行新技术新应用涉军信息内容监管职责。“行政机关收到检察建议后，已依法履行属地管理责任，督导涉案企业加强大模型安全评估，完善全流程内容安全管理体系。”

就在日前，徐汇检察院对整改情况开展了案件“回头看”。“我们发现涉案企业已对生成的违规图片进行溯源清理，同步上线了涉军规范示例34条、形象提示词28个等描述策略，并进一步细化管控规则，健全了AI生成合成内容审核机制，强化语料清洗与权威数据赋能，实现对相关提示词的有效识别与拦截。”张爱菁告诉记者，为推动实现从办理一案到治理一片，徐汇区目前正在对辖区内70余家备案企业开展综合网络安全检查，助力行业开展一次“全科体检”。

法律边界何在？产业、个人、司法均面临挑战

从AI生成涉军违法不良信息的个案窥视，如今，生成式AI的应用正遍及千行百业，带来更为广泛复杂的挑战。

此前，女演员温峥嵘就曾遭遇AI假冒自己虚假直播，起因是有人发现她曾“出现”在多个直播间，同一时段，妆造不同，产品不同，但说着同样的话。这些直播间影像均非温峥嵘本人出镜，而是有人利用人工智能进行“移花接木”。

安徽铜陵的一名女子则利用AI技术生成了“流浪汉进家”的图片，发给在外聚餐的丈夫，信以为真的丈夫立马报了警，民警到现场核实后，确认是利用AI整蛊引起的闹剧。

“从技术上来说，AI变声、AI换脸的门槛比较低，这是导致AI滥用的主要原因。”徐汇检察院普通犯罪检察部检察官龚笑婷坦言，这一新兴领域的确带来了许多新的问题。

“其主要问题在于衍生犯罪。”龚笑婷透露，在此前办理的一起案件中，就有企业通过在系统中的技术突破，生成淫秽物品从而在社会面上传播。“对于个人，可能也会因为AI变声、换脸，增加遭受电信诈骗的风险。而一旦这些数据被泄露或滥用，个人隐私也将受到严重侵犯。”

在知识产权领域，则同样面临问题。盛趣游戏知识产权总监郑慧指出，借助生成式人工智能技术产出的素材，可能无法被认定为“作品”，使企业面临作品无法进行商业化的难题，同时也面临被抄袭、盗用的风险。“AI可能无意识地‘借鉴’了他人享有著作权的内容，可能使企业在无主观侵权意图的情况下，因AI内容的‘相似性’陷入侵权纠纷，从而对企业产生负面影响。”

而对于司法机关而言，AI技术的快速迭代和内容生成的隐蔽性，给案件侦查、取证和定性也带来了前所未有的挑战。

“传统网络犯罪案件我们通过电子数据来鉴定，但这些新类型案件技术认定上会比较困难，因为AI生成的声音、图像甚至视频，往往难以通过传统鉴定手段辨别真伪，取证过程需要依赖专业技术支持，而相关鉴定标准尚在建设中。”徐汇检察院普通犯罪检察部主任顾伟介绍，包括在具体案件的责任划分中，比如针对因智驾引发的交通事故，由于涉及大模型提供者、车企、肇事者等多方主体，责任划分较为复杂。

此外，顾伟还提到，在AI医疗领域，还另外涉及到伦理挑战。“医疗AI需要大量的患者数据来训练和学习，这就涉及到患者的隐私保护问题。此外，如果因为AI误诊导致病情延误，责任归属问题也会变得非常复杂。”顾伟表示。

破解AI治理难题 需要形成协同治理大格局

作为三大先导产业之一，上海正在加速推动人工智能产业发展，持续落实中央“人工智能+”行动，推动落地“模塑申城”实施方案。统计数据显示，截至2024年，上海人工智能产业规模突破4500亿元，同比增长超过7.8%。今年一季度，上海规上人工智能产业规模超过1180亿元，同比增长29%，利润增长65%，成为拉动经济增长的新引擎。

▼ 检察官就案件展开讨论

上海市人工智能行业协会秘书长钟俊浩认为，在这样的发展背景下，司法机关正成为平衡创新与规范的“关键支撑”。“通过典型案例的裁判与司法解释的明确，司法机关可以为AI技术的应用划定清晰的法律边界与行为准则，尤其在数据安全、隐私保护、算法公平等方面树立裁判规则，为行业提供稳定、可预期的法治环境。”

钟俊浩建议，未来可以通过构建“分级分类治理体系”，提升治理精准度。“针对如生成合成、自动驾驶、医疗诊断等不同风险等级的AI应用场景，可以制定差异化的监管要求和司法审查标准，推动建立‘低风险促发展、中风险强监管、高风险严准入’的梯次治理模式，实现精准识别、精准施策。”

记者注意到，作为上海国际科创中心建设的重要承载区，目前，徐汇区正在通过高质效办案推进人工智能治罪与治理。“今年我院跨部门组建人工智能综合履职团队，对涉人工智能案件实行线索共享、一体研判，同时还出台了《服务保障人工智能产业发展三年行动方案》，部署18项责任项目。”徐汇检察院代理检察长孙军介绍。

“人工智能治理兼具复杂性和系统性，治理绝不是单靠一个部门的力量，更需要携手政府、企业、高校，共同构建开放协同的法治生态圈。”孙军表示，未来，徐汇区将围绕人工智能技术演进带来的新型法律挑战，高质效办好每一个涉人工智能新类型案件，孵化一批具有典型性、引领性的标杆案例，探索与确立适用人工智能发展的“规则之治”，并进一步完善跨区域人工智能治理机制，为全国贡献更多“徐汇样本”“上海经验”。