生成式AI合成内容鉴别安全实践

演讲人：李雪鹏　　单位名称：观安信息

01 生成式AI现状 STATE
02 法律法规政策 POLICY
03 风险 QUESTION
04 我们的行动 ACTION

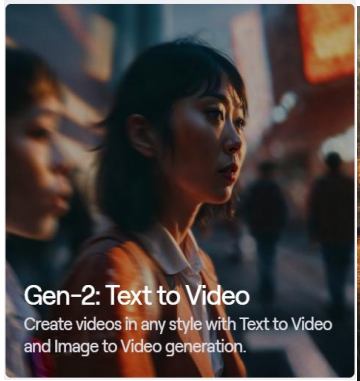
01生成式AI现状/生成式AI与传统AI对比

与传统AI的不同

生成式AI（Generative AI）是一种利用机器学习和人工智能技术生成新内容的系统。与传统的AI不同，生成式AI不仅能分析和理解现有数据，还能基于这些数据生成新的、未曾存在的内容。这些内容可以是文本、图像、音频、视频等各种形式。生成式AI的核心在于其能够创造性地生成具有一定创意和实用价值的新数据，从而拓宽了人工智能的应用领域。

现在所用技术

传统生成对抗网络（GANs）：由生成器和判别器两个对抗性网络组成，通过两者的对抗训练生成越来越逼真的数据
变分自编码器（VAEs）：过将数据编码为潜在变量，再从潜在变量生成新数据
自回归模型：Transformer、GPT等，通过逐步预测下一个数据点，自回归模型能够生成连贯的序列数据
扩散模型：过模拟数据生成过程中的噪声扩散和逆扩散过程，生成高质量的新数据



图片来自OpenAI、文心一格、Runway、StabilityAI、造梦日记、通义万相

01生成式AI现状/DeepFake热点事件

- 2024年 “三只羊”事件水落石出，AI雷军骂人的视频大量AI生成的真假难辨
- 2023年5月8日，包头警方发布了一起利用智能 AI 技术进行电信诈骗的案件
- 2023年2月17日，一则关于“杭州市政府将于3月1日取消机动车尾号限行政策”的消息也在网络疯传
- 2021年4月27日消息，荷兰议会外交事务委员会在视频会议软件Zoom上被AI换脸技术所骗
- 近期在直播领域，许多网友也发现多位带货直播博主都长了一张“明星脸”，神似明星杨幂、迪丽热巴、angelababy、佟丽娅等。
- 浙江温州市公安局通报过一起利用 AI 换脸技术进行诈骗的案件。

02法律法规政策/相关法律法规

2024年	《促进和规范数据跨境流动规定》《生成式人工智能服务安全基本要求》
2023年	《个人信息出境标准合同办法》《生成式人工智能服务管理暂行办法》《数字中国建设整体布局规划》《网络安全标准实践指南—生成式人工智能服务内容标识方法》《人工智能面向机器学习的数据标注规程》《信息安全技术 大数据服务安全能力要求》
2022年	《数据出境安全评估办法》《互联网信息服务深度合成管理规定》《工业和信息化领域数据安全管理办法(试行)》《信息安全技术 网络数据处理安全要求》《信息安全技术信息安全风险评估方法》
2021年	《数据安全法》《个人信息保护法》《关键信息基础设施安全保护条例》《网络安全审查办法》《互联网信息服务算法推荐管理规定》
2020年	《信息安全技术个人信息安全影响评估指南》《信息技术大数据 数据分类指南》《信息安全技术个人信息安全规范》
2019年	《互联网个人信息安全保护指南》《App违法违规收集使用个人信息行为认定方法》《信息安全技术数据安全能力成熟度模型》《信息安全技术 大数据安全管理指南》
2018年	《科学数据管理办法》《信息技术 数据质量评价指标》
2016年	《网络安全法》
2015年	《国家安全法》《促进大数据发展行动纲要》

质量参差不齐的AI生成内容病毒病毒式扩散，互联网似乎难逃“劣币驱逐良币”的宿命。

很多人都发现，抖音、快手、小红书等平台上的“抽象”内容越来越多，很多甚至已经违反了自然规律和常识，有些内容更是在道德和法律红线上反复横跳。——《面对AI生成的劣质内容，大平台正在反击》

赛博“照妖镜”：AI魑魅的标识与辨别

右图中人物的牙齿部分非常怪异，胸牌的颜色也露出了马脚（来自ClaudeAI鉴别工具）。

头部直播公司三只羊的“录音门”事件

有网友说：“你的脸在上传AI相机的那一刻，就不再属于你了。”

如：某手机的一键擦除、大名顶顶的DeepNude

原始图

分析图

04我们的行动/构建源于“攻防实践”的AIGC检测平台

应用场景

身份核验素材真实性校验

媒体内容真实性校验

产品功能

训练平台	文本鉴别算法	图像鉴别算法	原子能力开放
模型训练	文本困惑度	领域特征分离	文本鉴别
模型选用	文本规整度	细粒度分类	图像鉴别
数据增强	文本波动性	多模型融合	视频鉴别

攻击靶场

攻击手段 → 模型动态更新 ← 防守策略

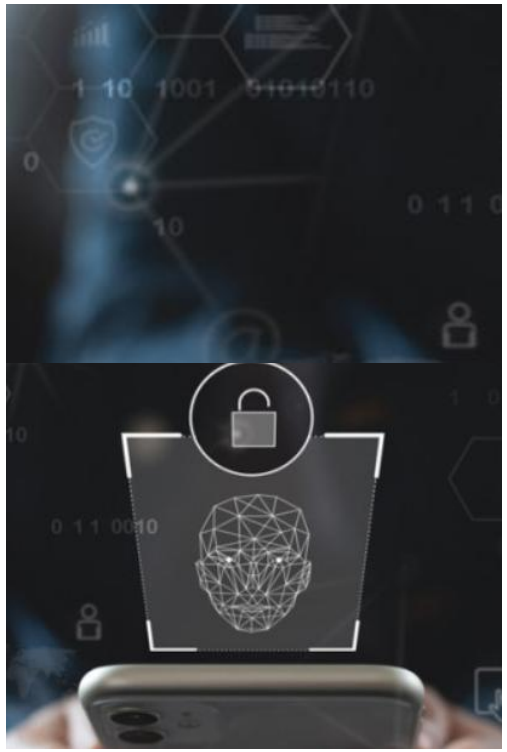
海量数据

海量精准标注的真实行业数据、大规模多种类动态更新迭代的AI合成数据

04我们的行动/应用场景

合成检测和活体检测能力是以模块化方式提供的，方便其他平台系统以接口的方式调用。也可以提供云端部署和托管选项的方式，用户无需自行配置和维护硬件设备，只需通过网络访问相应的服务接口即可使用合成检测功能，大大减少了人工维护成本。

基于合成检测和活体检测的人脸识别技术成果可推广的领域



通信行业	银行证券	金融保险	民生社保	在线教育	共享服务	新闻媒体
反诈检测	手机app身份验证	远程营销身份确认智能出险	免排队线上办理	远程登录身份验证	免密登录	实名认证防造谣
手机app免密登录	刷脸转账远程验证		手机端身份验证	在线考试防作弊		反欺诈案例宣传
远程业务办理身份确认	远程支付活体验证		远程操作社保报销			

目前，实名信息自动稽核平台已在某省电信投入使用，帮助线上渠道的发卡业务中实名信息进行二次稽核，总计**涵盖7类场景，40+模型**，包含多种渠道下的证件稽核、活体照片、免冠照片、活体视频等稽核场景，保障企业运营安全。

上线之后6个月内累计发现20000+不合规订单、400+疑似涉诈的订单，有效补足了业务校验规则缺失部分，提高系统业务合规性。

AI合成人脸样例

DeepFake换脸

AI生成人脸

实名信息拍照不规范典型样例

闭眼

不清

不全

戴帽

光斑

光线暗

口罩遮挡

现场照片为
一寸照照片

PS照片

身份证不清晰

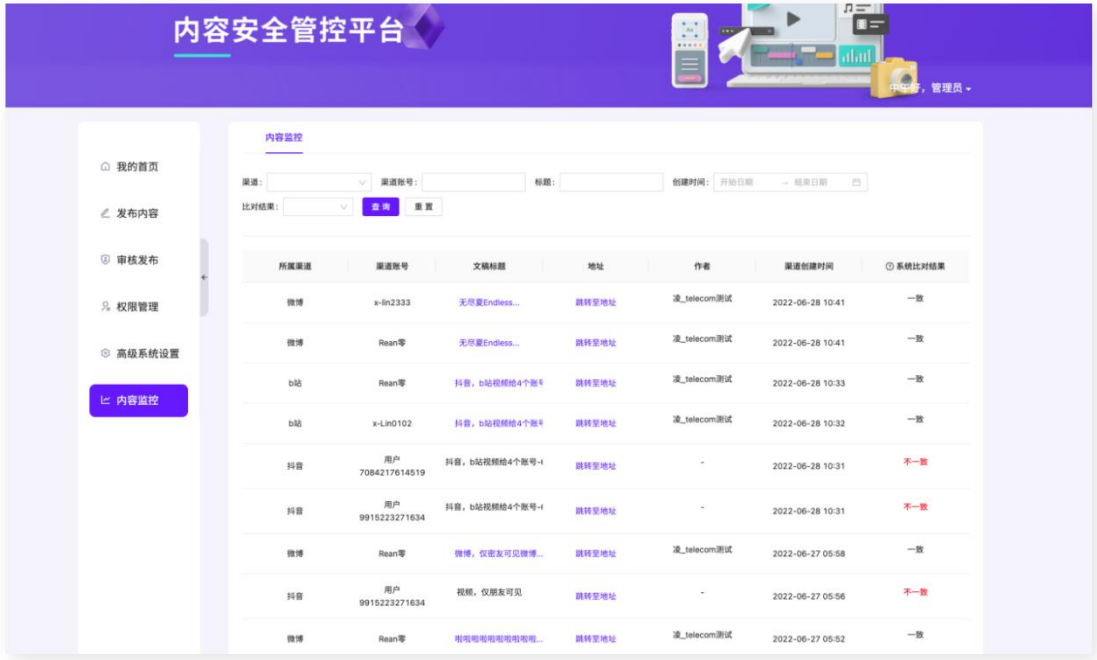
视频拼接卡顿

视频PS

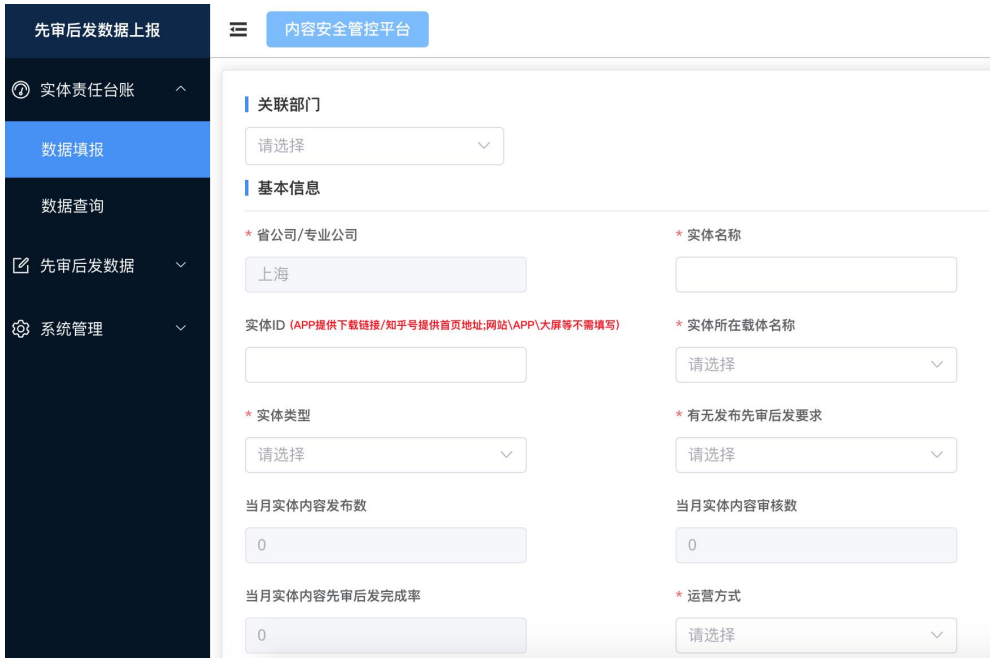
04我们的行动/数字内容风控-媒体内容先审后发

目前，**平台已在某省电信投入使用**，覆盖全部业务部门，业务角色有发布人员、审核人员以及部门管理员。提升自上而下的监管效率，大大降低信息内容安全风险，经审批后的信息不可篡改（可稽核监控），整个流程可追溯，为事后准确追责与问题快速定位提供真实数据支撑。

内容安全管控平台



先审后发数据上报平台



04我们的行动/能力体验

观安信息基于多年大数据以及人工智能领域积累，研发**观安白泽AI鉴别系统**，可以鉴别各类由Deepfakes生成的虚假内容，包括文本内容、图片以及视频。

（扫描二维码即可体验观安信息合成内容检测能力）

致力于“AI向善发展”

THANK YOU!