金融AIGC音视频反欺诈白皮书

2024.12

版权说明

编写组

主编： 李肇宁

副主编： 钱菲、陈树华、田天

参编人员：

王光中、赵晗、艾国、高峰、魏恪、 王继科 、 史博、宋文利、李煜明、刘荔园、萧子豪、刘汉鲁、孙空军、杨金威

参编单位：

交通银行股份有限公司、北京顶象技术有限公司、北京瑞莱智慧科技有限公司

序

早在 2018 年，习近平总书记就强调要未雨绸缪，加强战略研判，确保人工 智能安全、可靠、可控。此后，习近平主席又在多个国际场合倡议“不断提升 人工智能技术的安全性、可靠性、可控性、公平性”“引领全球人工智能健康 发展”[1]。在此背景下，我国陆续出台了一系列法律法规与政策文件，以加强 AI 的安全监管和规范应用。2024 年 7 月，二十届三中全会通过的《中共中央关 于进一步全面深化改革、推进中国式现代化的决定》中，特别强调了“完善生 成式人工智能发展和管理机制。”“加强网络安全体制建设，建立人工智能安 全监管制度。”[2]这是党中央统筹发展与安全，积极应对人工智能安全风险作 出的重要部署。为此，国内发布了包括《国家新一代人工智能标准体系建设指 南》、《生成式人工智能服务管理暂行办法》和《关于依法惩治网络暴力违法 犯罪的指导意见》等多项政策，明确对利用深度合成技术发布违法信息的行为 从重处罚。

在金融领域，基于人工智能的 AIGC 技术的普及带来了显著的创新潜力，但同时也给金融机构的业务安全、客户信任以及系统稳定性构成了新的挑战。特别是音视频领域的 AIGC 欺诈手段，已经成为金融机构必须面对的重要风险之一。这些欺诈行为不仅损害了金融机构的声誉和利益，更对广大客户的财产安全构成了严重威胁。

AI 治理攸关全人类命运， 必须采取切实有效的措施， 贯彻人工智能安全理念，防范 AIGC 欺诈，保障金融业务安全。一方面，要加强技术研发和创新，提升 AIGC 技术的安全性和可控性。通过加强算法研究、优化模型设计、提高数据质量等手段，不断提升 AIGC 技术的准确性和稳定性，减少其被恶意利用的风险。另一方面，要加强监管和治理，建立健全人工智能安全监管制度。通过完善法律法规、加大执法力度、提高监管效能等手段，确保人工智能技术在金融领域的应用符合法律法规要求，保障金融业务的合规性和安全性。

基于此，交通银行、顶象技术、瑞莱智慧联合编写了《金融 AIGC 音视频反欺诈白皮书》，通过详实的数据、典型的案例和前瞻性的技术分析，系统介绍 AIGC 带来的欺诈风险，深入剖析金融机构面临的 AIGC 音视频风险挑战，并提出 AIGC 音视频反欺诈方案、技术实现路径及相关倡议，以期为金融机构提升 AIGC 欺诈识别和防范能力提供有益参考。

相信通过强化合规体系建设， 加强反欺诈技术创新， 构建全链条健康生态， 守正创新携手共进，必将推动人工智能的健康发展，赋能金融高质量发展。

交通银行副行长兼首席信息官：

目录

序 1

第一章 AIGC 带来的音视频欺诈风险 5

1.1 AIGC 驱动音视频技术创新的同时带来新风险 5

1.1.1 图像和视频合成技术的发展 5

1.1.2 音频合成技术的发展 6

1.2 AIGC 带来的“换脸”风险 6

1.2.1 AIGC“换脸”的技术原理 6

1.2.2 AIGC“换脸”的主要应用场景 6

1.2.3 AIGC“换脸”带来的安全挑战 7

1.3 AIGC 带来的“拟声”风险 7

1.3.1 AIGC“拟声”的技术原理 7

1.3.2 AIGC“拟声”的主要应用场景 8

1.3.3 AIGC“拟声”带来的安全挑战 9

1.4 AIGC“换脸”“拟声”风险的特征 9

1.4.1 生成内容的高仿真性 10

1.4.2 内容生成的低成本和高效率 10

1.4.3 难以溯源的隐匿性 10

1.4.4 跨模态内容生成与融合 10

第二章 AIGC 音视频欺诈典型攻击方法 12

2.1 AIGC“换脸”攻击分析 12

2.1.1 AIGC“换脸”攻击目标 12

2.1.2 AIGC“换脸”攻击过程 13

2.1.3 AIGC“换脸”攻击技术 14

2.2 AIGC“拟声”攻击分析 15

2.2.1 AIGC“拟声”攻击目标 15

2.2.2 AIGC“拟声”攻击过程 15

2.2.3 AIGC“拟声”攻击技术 16

第三章 AIGC 音视频欺诈对金融业务的影响 17

3.1 增加金融业务风险 17

3.2 给黑灰产攻击提供新手段 17

3.3 为防御带来新挑战 18

3.4 对金融反欺诈提出新要求 19

第四章 AIGC 音视频反欺诈方案 20

4.1 构建全面防御体系 20
4.2 技术解决思路 21
4.2.1 多模态 AIGC 音视频欺诈的检测技术 21
4.2.2 多模态 AIGC 音视频欺诈的鉴定技术 23
4.2.3 AIGC 特征的欺诈团伙识别技术 24

4.2.4 融合 AIGC 欺诈的多模态智能决策引擎技术 26
4.3 从业人员能力的提升 28
4.4 管理体系的提升 29
4.5 法律法规护航 30
4.5.1 针对 AI 滥用的法规 30
4.5.2 针对违法者的惩罚 31
第五章 AIGC 音视频反欺诈技术实现 32
5.1 AIGC 音频伪造检测 32
5.1.1 语音伪造线索 32
5.1.2 线索建模方式 33
5.2 AIGC 图像伪造检测 34
5.2.1 图像伪造线索 34
5.2.2 线索建模方式 35
5.3 AIGC 视频伪造检测 36
5.3.1 视频伪造线索 36
5.3.2 线索建模方式 38
5.4 AIGC 欺诈鉴定技术 38
5.4.1 被动式溯源 38
5.4.2 主动式溯源 39
5.5 基于知识图谱的特征关联分析 40
5.5.1 基于 AIGC 特征的关系建立 41
5.5.2 发现与识别团伙欺诈 41
5.5.3 提升反欺诈的能力 42
5.6 融合反 AIGC 欺诈计算引擎的处理系统 42
5.6.1 数据采集与预处理 42
5.6.2 特征与规则 43
5.6.3 智能决策引擎与风险评估 43
5.6.4 实时响应与行为拦截 43
5.6.5 业务价值及优势 43
第六章 典型业务场景 45
6.1 远程音视频反欺诈 45
6.1.1 背景 45
6.1.2 风险分析 45
6.1.3 解决方案 45
6.1.4 实施效果 46
6.2 人脸识别身份认证反欺诈 46

6.2.1 背景 46
6.2.2 风险分析 46
6.2.3 解决方案 47
6.2.4 实施效果 48
6.3 伪造人脸考勤反欺诈 48
6.3.1 背景 48
6.3.2 风险分析 48

6.3.3 解决方案 48
6.3.4 实施效果 49
6.4 虚假视频聊天反欺诈 49
6.4.1 背景 49
6.4.2 风险分析 49
6.4.3 解决方案 50
6.4.4 实施效果 50
第七章 展望与倡议 51
7.1 未来技术挑战 51
7.2 相关倡议 51
7.2.1 健全合规体系 52
7.2.2 创新发展技术 52
7.2.3 构建健康生态 53
后记 55
参考文献 56

第一章 AIGC 带来的音视频欺诈风险

生成式人工智能（AIGC，Artificial Intelligence Generated Content）技术的迅猛发展，推动了内容生成领域的广泛应用，涵盖了文本、图像、音频、视频等多模态内容生成，为娱乐、教育、营销及各行各业的应用带来了前所未有的创新。然而，AIGC 的应用与普及也带来了新的风险挑战，亟需多方监管、加强技术检测与防范措施，确保其在商业应用的安全与透明性，同时加强用户教育以提升风险防范意识。

1.1 AIGC 驱动音视频技术创新的同时带来新风险

AIGC 已逐步渗透至各个应用场景中。其背后强大的技术支撑包括图像和视频的生成对抗网络（GAN）、扩散模型（Diffusion Model）、神经辐射场（NeRF）等一系列深度学习技术，以及音频合成中的文本到语音（TTS）和语音转换（VC）等技术。这些技术的进步不仅显著提升了 AIGC 内容的质量和生成效率，也带来了在娱乐、社交、金融等多个行业的广泛应用及新的风险。

1.1.1 图像和视频合成技术的发展

生成对抗网络（GAN）。生成对抗网络（GAN）是 AIGC 技术的基础之一，它通过生成器和判别器的对抗训练，不断优化生成内容的质量。生成器负责创造出新的图像或视频内容，而判别器则尝试辨别生成内容是否与真实内容相似，从而在不断对抗的过程中提升生成内容的真实性。GAN 技术已经实现了高度逼真的图像和视频生成效果，使得深度伪造成为可能。这一技术的应用场景包括人脸替换、虚拟化身创建等，但同时也为伪造视频的生成提供了可能。

扩散模型（Diffusion Model）。随着深度学习算法的进步，扩散模型逐渐成为 AI 视频伪造领域的新兴主流技术路径之一。扩散模型通过在噪声中不断增加与还原信号的过程，能够生成非常逼真的图像和视频序列。扩散模型不仅在生成效果上比 GAN 更为出色，且生成过程更为稳定，其在细节处理、光影效果等方面的表现尤为显著。这使得扩散模型在高保真视频和复杂场景的伪造方面具有巨大的潜力。

神经辐射场（NeRF）。神经辐射场（NeRF）技术的出现为 3D 重建与人脸伪造提供了新的方向。NeRF 通过学习光线在 3D 空间中的辐射强度分布，能够实现复杂的 3D 重建和高保真的人脸伪造。这种技术能够将 2D 图像数据重构为 3D 场景，并生成逼真的视觉效果，使得人脸伪造的真实感更高。与 GAN 和扩散模型相比，NeRF 更适用于 3D 场景的模拟与重建，因此其在元宇宙、虚拟现实等领域也具有广阔的应用前景。

当前，以 GAN、Diffusion 和 NeRF 为基础的技术路线在图像和视频伪造领域呈现出三足鼎立的趋势。这三种技术各有优势，分别在 2D 人脸伪造、复杂视频生成、3D 人脸重建等方面各显其长。这些技术的不断演进，使得 AI 视频伪造的质量、速度和逼真度不断提升，带来了更广泛的应用可能性。

1.1.2 音频合成技术的发展

文本到语音（TTS）。文本到语音（Text-to-Speech，TTS）技术通过将文本输入转化为自然语音，实现了较高质量的语音生成。这一技术的核心在于如何使合成语音听起来自然、流畅，并具有一定的情感表达能力。当前的 TTS 技术可以在短时间内生成高保真的语音，使得虚拟助手、虚拟主播等应用能够轻松模仿真人的语音风格。

语音转换（VC）。语音转换（Voice Conversion，VC）技术是另一种关键的音频伪造技术，通过将源语音的特定属性（如音色、语调）转换为目标语音的特征，从而生成与目标人物相似的语音内容。不同于 TTS，VC 技术在保留语音内容的前提下，能够改变语音的特征，使其听起来更接近目标人物。基于深度学习的 VC 技术相比早期的统计建模方法，生成效果显著提升，能够更真实地模拟目标语音风格。

风格迁移和语音大模型。在语音伪造领域，风格迁移技术进一步提升了合成语音的自然度和真实性。通过模拟目标语音的说话风格和情绪特征，风格迁移弥补了传统语音合成在情感表现上的不足。同时，语音大模型的出现进一步提高了语音合成的质量和效率。如今，仅需少量的音频样本便可生成高质量的语音合成内容，这使得高精度、低成本的语音伪造成为现实。

1.2 AIGC 带来的“换脸”风险

1.2.1 AIGC“换脸”的技术原理

AIGC“换脸”技术，是指利用 AIGC 技术，通过对目标视频或图像中的某个人的面部进行替换，将其变为另一个人的面部。此技术依托于深度学习框架，尤其是生成对抗网络（GAN）和大型预训练模型，通过大量人脸数据进行训练，以生成高度逼真的“换脸”效果。GAN 由生成器（Generator）和判别器（Discriminator）构成，生成器负责生成与真实数据难以区分的“假数据”，而判别器则负责判断生成的图像真假，二者不断对抗，优化生成效果，最终生成逼真的人脸替换效果。

通过 GAN 和其他模型的协同，AIGC“换脸”技术能够学习到人脸的细微表情、光线反射、纹理细节等因素，在面部表情变化、嘴唇与声音同步、光影调整等方面取得了极高的真实度。此外，AIGC“换脸”技术的生成过程也因其高度自动化而具备较强的泛化能力，无需过多人工干预便可以实现逼真且多样化的面部替换效果。

1.2.2 AIGC“换脸”的主要应用场景

影视与娱乐行业。AIGC“换脸”技术在影视制作中的应用广泛。例如，可以用明星的面孔替换替身演员的面孔，使表演更加真实且减少重复拍摄需求。

此类技术也被应用于影片复原或重拍，将已故演员的形象复现到影片中。此外，AIGC“换脸”技术在虚拟主播、数字偶像等新兴娱乐领域中同样受到关注。

社交媒体和创作。 在社交媒体上，“换脸”特效让用户能够体验角色扮演 的乐趣，迅速成为热门潮流。通过 AIGC“换脸”技术，用户可以在短时间内生 成内容，便捷地分享具有高度真实感的“换脸”视频。此技术不仅为用户提供 了极大的创作空间，也为个性化内容的生成和传播提供了可能性。

虚拟现实与增强现实应用。 AIGC“换脸”技术在虚拟现实（VR）和增强现实（AR）领域也具有重要作用。例如，利用“换脸”技术可以使用户在虚拟场景中拥有不同的面孔，从而进一步提升沉浸感。无论是游戏角色的面貌定制，还是 AR 社交平台上的角色扮演，这类应用都因 AIGC“换脸”技术而变得更具吸引力和互动性。

1.2.3 AIGC“换脸”带来的安全挑战

尽管 AIGC“换脸”技术在多个领域展现出潜力，但其也带来了安全和道德上的挑战。黑灰产可能利用此类技术在未经授权的情况下非法使用他人肖像，甚至对当事人形象进行恶意篡改或丑化，存在侵犯肖像权、名誉权及隐私权的风险。

2024 年 1 月，美国知名歌手泰勒 ·斯威夫特的伪造“不雅照片”在 Facebook 等社交平台广泛传播，累计浏览量超过千万。尽管最初传播该照片的账号已被封禁，但照片的扩散仍未彻底遏制，严重侵犯了泰勒 ·斯威夫特的个 人权益。

2023 年 5 月 23 日，包头警方公布了一起利用 AI 实施电信诈骗的典型案例， 福州市某科技公司法人代表郭先生在短短 10 分钟内被骗走 430 万元人民币。

2024 年 2 月，一黑灰产通过“换脸”技术伪造跨国公司高层身份，参与视 频会议指挥分公司向指定账户汇款，成功骗取 2 亿港元。

利用 AIGC“换脸”技术带来的风险主要在身份伪造与深度伪造、隐私泄露和滥用、社会信任的破坏等三个方面。随着实施此类犯罪的技术门槛逐步降低，并预计将持续上升。

身份伪造与深度伪造。 AIGC“换脸”技术的真实性使其成为一种极具隐患的身份伪造手段。

隐私泄露和滥用。 利用 AIGC“换脸”技术生成的假视频、假照片很可能在未经授权的情况下泄露他人隐私，甚至被用于恶意传播不实信息。这不仅侵犯了隐私权，还可能对受害人造成名誉损害。

社会信任的破坏。AIGC“换脸”技术的大范围应用可能削弱公众对视频和图像真实性的信任。例如，普通用户难以区分真假视频，进而对信息的真实性产生怀疑，甚至影响到司法调查、新闻报道等领域的公信力。

1.3 AIGC 带来的“拟声”风险

1.3.1 AIGC“拟声”的技术原理