WAIC2023 科技绘蓝图 标准引方向 合合信息携手中国信通院助力AI向善

多方主体共同发起《AIGC可信倡议》

开年以来，AIGC（生成式人工智能）技术火爆全球，人们对技术革新带来的美好生活倍感期待的同时，也忧心于频发的高科技诈骗事件。世界人工智能大会（简称WAIC）素有“科技风向标、产业加速器”之称，如何有效破解信息安全难题、构筑可信AI生态体系，愈发凸显出其紧迫性和重要性。可信AI也成为WAIC重点关注的议题之一。

面对AI引发的“信任焦虑”，大会期间，中国信息通信研究院（简称中国信通院）主办了“聚焦·大模型时代AIGC新浪潮”论坛，围绕“多模态基础大模型的可信AI”这一热门议题开设专场分享。为进一步推动可信人工智能发展，《AIGC可信倡议书》（以下简称“倡议”）面向全球正式发布。上海人工智能实验室、中国通信院、武汉大学、蚂蚁集团、百度、OPPO、合合信息等共同出席本次倡议发布仪式。

现阶段，用可信AI为技术创新和产业落地营造规范、稳健的发展环境已成为普遍的共识。而要想将共识转化落地，离不开“硬科技”的支撑。论坛分享中，合合信息面向公众展示了多项图像内容安全保护技术，涉及AI图像篡改检测、生成式图像鉴别、个人信息保护等多个领域，为业界展现了AI在图像领域可信化发展的多重可能。

AI图像篡改检测技术向截图造假“亮剑”

“P图”是人们工作、生活中常用的图片修饰手段，但“万物皆可P”的想法也导致了许多犯罪行为的发生，比如通过随意修改他人证件照上的肖像、个人信息，或着票据上的金额、时间，来误导他人，达到非法获利的目的。

在普遍“年轻”的AI企业里，有着十余年深耕经验的合合信息是一名“老兵”，通过智能文字识别及商业大数据技术，为全球超过200个国家和地区的上亿用户提供智能文字识别产品及AI服务。在深厚研发实力的支持下，合合信息不断为AI可信化落地带来多元化的技术赋能工具，助力行业健康发展。

2022年世界人工智能大会上，合合信息“PS篡改检测”技术对证件图片篡改痕迹进行了“像素级”的识别，引起广泛的关注。该技术基于深度学习的图像篡改检测技术及相关系统，通过学习图像被篡改后统计特征的变化，智能捕捉图像在篡改过程中留下的细微痕迹，并以热力图的形式展示图像区域篡改地点，覆盖身份证、护照等多种证照识别类目，已应用于银行、保险等领域多家机构。

今年，该技术升级为能够识别截图篡改痕迹的“AI图像篡改检测”技术。截图中通常包含了转账记录、隐私聊天记录、商业机密等重要信息，若遭受恶意篡改或泄漏，个人及企业可能遭受更大的财产损失。于普罗大众而言，这项技术将成为戳穿电信网络诈骗谎言的实用“武器”。

合合信息AI图像篡改检测技术可检测包括转账记录、交易记录、聊天记录等多种截图，无论是从原图中“抠下”关键要素后移动“粘贴”至另一处的“复制移动”图片篡改手段，还是“擦除”“重打印”等方式，图像篡改检测技术均可“慧眼”识假。

合合信息智能创新事业部总经理唐琪介绍AI图像安全相关技术成果

与证照篡改检测相比，截图检测难度更大。从成像角度来看，截图的背景没有纹路和底色，整个截图没有光照差异。证件篡改识别尚可通过拍照时产生的成像差异进行篡改痕迹判断，而截图则没有这些“信息”。

此外，现有的视觉模型通常难以充分发掘原始图像和篡改图像的细粒度差异特征，因此难以实现令人满意的准确率。合合信息提出了一种基于HRNet的编码器－解码器结构的图像真实性鉴别模型，能够捕捉到细粒度的视觉差异，达到高精度鉴别效果。

生成式图像鉴别用AI护航AI发展

除了图像篡改外，近期以“AI换脸”诈骗为代表的生成式图像欺诈手段也是保障AI图像内容安全所需要面对的难题。本次大会上，合合信息发布了生成式图像鉴别技术，帮助个人及机构识别判断AI图片原始属性，规避可能存在的欺诈、伦理等方面的风险。

生成式图像鉴别的难点主要分为两点。首先，生成出来的图像场景繁多，无法穷举，机器不能通过一一细分项目来训练模型、解决问题；再者，有些生成图和真实图片的相似度过高，很贴近于人类的判断，对于机器而言，真伪判定只会更难。

合合信息生成式图像鉴别技术的优势在于避开了常规的穷举检测思路，不追求持续扩容的场景细分来实现图像是否是生成式的检测判断，而是基于空域与频域关系建模，能够在不用穷举图片的情况下，利用多维度特征来分辨真实图片和生成式图片的细微差异。

生成式图像检测技术在反诈骗、版权保护等领域的应用空间十分广泛。例如在金融行业，不法分子可能会利用AI合成技术对线上资金进行盗刷，威胁公民财产安全。本项技术可通过对支付环节的干预，防止资金盗刷；在文化传播行业，某些图片供给方使用软件自动生成图片，故意隐瞒其来源并售卖给第三方，导致了系列侵权问题。生成式图像检测技术可在一定程度上降低这些事件发生的概率。

OCR对抗攻击技术信息安全“定向加密”

人们往往会出于生活、工作需要，拍摄自己的相关证件、文件并发送给第三方，这些图片上承载的个人信息可能被不法分子使用OCR技术识别提取并泄露。为满足个人、企业业务的文件资料保密需求，合合信息在论坛上展示了一个给个人图片信息定向“加锁”的创新项目。

据悉，该项技术名为OCR对抗攻击技术，通过场景文本或者文档内文本进行扰动的方式，实现对中文、英文、数字等关键信息的内容进行“攻击”的目标，防止第三方通过OCR系统读取并保存图像中所有的文字内容。这项技术不影响肉眼观看与判断，既可以满足生活、工作场景中的信息传输需要，也可以降低数据泄露的风险。

权威机构与企业携手建立行业规范

▲合合信息AI图像篡改检测技术应用效果展示

图像内容安全的重要性与日俱增。在这场隐形的安全保卫战中，除了技术之外，针对可信AI产品及服务的标准规范的出台，将有助于AI图像行业整体向合规方向发展。

从2015年7月国务院出台《关于积极推进“互联网+”行动的指导意见》首次将人工智能纳入重点任务，到《“十四五”数字经济发展规划》对人工智能产业的战略布局，中国在人工智能领域一直秉持审慎包容、鼓励创新的态度，既肯定其在经济社会发展中的重要价值，也一直在持续推进人工智能伦理规范和法律规范等方面的发展。

为贯彻落实《中华人民共和国网络安全法》《生成式人工智能服务管理办法（征求意见稿）》等文件中对于AI服务的规范性要求，系统性建立图像内容安全行业发展秩序，中国信通院牵头启动了《文档图像篡改检测标准》制定工作，合合信息、中国图象图形学学会、中国科学技术大学等科技创新企业及知名学术机构联合编制。

《文档图像篡改检测标准》将为文档图像内容安全提供可靠保障，助力新时代AI安全体系建设。作为牵头方，中国信通院表示，《文档图像篡改检测标准》将基于产业现状，围绕“细粒度”视觉差异伪造图像鉴别、生成式图像判别、文档图像完整性保护等行业焦点议题，凝聚共识，以期为行业提供有效指引。

合合信息智能创新事业部总经理唐琪在论坛分享中表示，《文档图像篡改检测标准》制定项目的启动，是AI图像内容安全体系建设之路上的重要的里程碑。期待更多研究机构、企业主体参与进来，共同推动AI服务规范性的整体提升和行业发展秩序的建立，助力科技向上而行，向善发展。