视频通话里的“熟人”也可能是假的？

鉴伪师揭秘AI造假套路

AI生成的照片获得某摄影比赛一等奖，让AI造假问题引发新一轮关注。比起存在明显破绽的AI照片，还有一些AI生成内容更加高明，潜藏在我们身边也更加危险。面对AI造假问题，有专业人士正通过技术手段与之博弈，在这场不断升级的信息攻防战里守护我们的安全。

算法研究员演示AI鉴伪系统如何对AI造假内容进行检测

屏幕外普通人 屏幕里变成“马斯克”


AI技术可以做到实时换脸，还能打视频电话。

指尖敲击键盘，电脑屏幕上代码不断滚动，各种文字、音频、图片和视频被录入系统，去当作“喂”给AI的原料。中科睿鉴公司算法研究员何覃一边关注着模型的训练过程，一边调整各种参数。

他正在训练的，是一款AI鉴伪模型。中科睿鉴主要做的就是研发AI鉴伪技术及产品，以应对可能存在的各种AI造假问题。从某种意义上说，何覃等人算得上是“AI鉴伪工程师”。

“现在AI技术已经非常发达，有些通过AI生成的内容凭肉眼已经很难分辨。如果有不法分子刻意利用AI来造假，有可能会引发多种问题和危害。”何覃介绍，像前段时间经常被讨论的“AI造假仅退款”问题，有消费者在购买产品并收到实物后，会拍摄一张照片，用AI把照片里的实物加上各种瑕疵，以此作为证据来“维权”并申请退款，这种现象也让一些商家头疼不已。

还有的AI造假，会给老百姓的财产安全带来风险。比如有不法分子利用换脸换声技术模仿明星，在平台发布假视频，塑造单身人设，专门对中老年人实施情感诱导，待时机成熟之后，再用各种理由骗取老年人的钱财。

何覃表示，有些假视频，是先拍一段真视频，后期再用AI技术进行换脸。而目前的AI技术，已经可以做到在实时通话中实现换脸。

公司的另一位工程师董宏雨现场演示了实时换脸技术。在一个软件里，有各种人物照片，随意点击一张，人脸就能实时换成对应的人。换脸成功之后，董宏雨甚至还能给记者拨来视频通话，屏幕里的形象很像外国企业家马斯克的样子。

“小孩和老人分辨能力弱，如果有人利用他们身边亲戚朋友的脸进行诈骗，后果不堪设想。”何覃表示，他和同事研发的AI鉴伪技术，就是为了在人的肉眼无法分辨AI内容的时候，提供一个机器识别的路径。记者将换脸假视频导入鉴伪软件之后，系统果然给出了“假”的判定。

摸索挖掘痕迹 想鉴假得懂造假

国家反诈中心App里搭载了AI内容鉴定功能。

何覃和同事们研发的鉴伪软件，采用的也是AI技术，相当于是在“用AI打败AI”。何覃表示，一开始训练模型的时候，只能通过做标记、灌数据等方式让模型自我学习，并做出判断。但如何向用户解释模型做出这些判断的依据是什么、判定AI造假的证据是什么，是该技术领域的一大挑战。

“后来经过我们的摸索和挖掘，逐渐找到了一些AI造假的共性痕迹，比如AI目前是没法模拟自然光的。”何覃表示，在现实世界中，光照射到物体表面时，会由于材质、表面与光照夹角等因素引起光照强度、空间分布的规律变化，图像中体现为特定的亮度梯度，提取这些信息可以计算出光源方向等光照特征。这种反射现象是肉眼难以察觉的，但在自然拍摄的视频和图片里客观存在。“AI造假合成的时候，它通常是靠算法模拟生成的，只是模拟了视觉逼真效果，没有模拟出遵循光学或物理学原理的现象。”

除了光源问题，一部拍摄设备拍出的照片和视频，因为电子设备元器件的特性，还会留下“物理噪声”，它并非人耳能听到的那种噪声，而是一种信号，可以通过技术手段提取出来，用视觉化的方式将它展现出来。AI生成或经过AI处理的图片和视频，在物理噪声方面与正常的图片和视频是不一样的。何覃表示，AI产出的结果，想要同时欺骗真人和机器是很难的，想要骗过人眼，就会在机器上留下痕迹，而如果只是为了骗过机器，产出的结果又很可能被人眼轻易识破。

为了能找到这些“痕迹”，技术人员要去对模型产出的结果进行研究和归纳，也需要不断去研究市面上先进的AI生成大模型，并尝试复现和理解它生成AI内容的过程和算法原理。“就像古玩行业，你想要鉴假，你得搞清楚它是怎么造假的，用了什么方式去‘做旧’。AI鉴伪也是一样，要知己知彼，才能知道鉴定时从哪里入手。”

想要防御风险 实时鉴伪很重要

做AI鉴伪行业多年，何覃最大的感受是，AI技术的更新速度实在太快了，造假也变得越来越容易，这就给“AI鉴伪工程师”提出了一个严峻的挑战。何覃表示，目前他和同事们在花精力研究的问题，就是AI鉴伪模型的“泛化性”。一方面，它能否应用在不同的场景，针对不同AI造假内容都能做出正确的判断。另一方面，它能否在一定程度上覆盖未来可能出现的“新型造假技术”。

在AI技术不断迭代的背景下，普通人该如何避免自己的信息被不法分子利用，拿去AI造假？在何覃看来，想做到这一点几乎是不可能的。“人毕竟有社交需求，总会在网络上留下照片等个人信息。如果为了防止身份被AI盗用，就刻意把自己封闭起来，不在网络上留下痕迹，是不现实的。”也正因如此，通过技术手段去鉴别AI信息并及时给予预警，在何覃看来是一个可行性更高的对抗AI造假的方案。“我们希望AI鉴伪技术能被当作一个日常工具，普通人也可以随取随用，保护自己的信息安全。”

何覃的设想，已经逐步走进现实。在国家反诈中心和学习强国App中，就搭载了AI内容鉴定的功能，采用的就是他们研发的技术。记者用几张真实照片和AI生成照片分别进行了测试，系统都能准确分辨。但记者也发现了一个问题，这个检测系统必须主动上传图片、视频等内容，才能给出鉴定结果。如果是实时换脸视频通话，系统是没法做到实时检测的。

“这也是我们在考虑的一个问题，有些假内容、假信息是通过实时方式传播的，如果当时没有发现，事后再想去检测，很可能已经晚了。”在何覃看来，更加有效的AI鉴伪方式，是在个人使用的手机、电脑等设备上，搭载可以随时调用的AI鉴伪技术。一旦出现AI假信息，用户可以实时检测，或者由系统发出报警。

在AI技术如此发达的今天，何覃建议，普通人也应该适当学习一些AI知识，提高认知，至少应该了解AI能做到什么，AI能用哪些方式造假，才能加以提防。

据《北京晚报》