上图：下载国家反诈中心 App 对老年人来说是一个较为管用的防诈骗手段。

这种方式去反哺 AI，进一步改进；哪些可能又是一些新的特征，下次要把它加入到可以识别的范围内来，通过这样不断的训练，让 AI 越来越智能。

在鄢晓玲看来，硬件、算法和数据，对于 AI 来说，是三个核心要素。“硬件，可以把它理解为是教室，它是 AI 能够去学习的物理条件。算法是老师，它给 AI 模型的训练制定规则，解决怎么学习的问题；数据则是教材，它决定了 AI 模型会学到点什么，也就是学什么的问题。由此不难看出，算法和数据非常重要，如果模型设计中带有偏见、不公、歧视或者某些缺陷，又或者数据中包含了非常多‘杂质’，那么培育出来的 AI 就有可能‘失控’。”

举个简单的例子，在涉政领域，对于国家主权、种族歧视，还有各种分裂者的一些观点，训练者会把有这种有诱导性的问题抛给模型，看模型怎么回答。然后通过人工干预，让 AI 建立起边界感，“不能一本正经地胡说八道”。“就像在学校里学习的孩子，毕业的时候要经过一个基本的考试。面向社会公众使用的 AI 模型，同样应该经过一些基线评测。”鄢晓玲说，他们正在向相关部门提供一些方案，力争能够早日达成业界共识。

萧子豪则表示，瑞莱智慧一直在研究深度合成技术的自动化检测，而他们主要的鉴别方式确实可以用“魔法打败魔法”来概括，也就是用 AI 去打败 AI。

常用的方法包括基于伪造内容数据集完成对模型检测器的训练、基于帧间不一致性实现对伪造内容的判别等，这些方法在公开数据集中均能达到 99.9% 的准确率，在产业实践中的检测准确率也已达业内顶尖水平。通过这些技术手段，可以准确地在某些金融机构的实名认证环节中拦截深度伪造攻击、监测到互联网中的明星 AI 形象带货直播。

目前防范的难度主要在于 AI 技术不断自我优化、升级迭代，深度合成内容质量不断提升。换声的自然度以及换脸视频的逼真度、流畅度都在不断提高，逐渐模糊了真实和虚假的边界，传统鉴别方式越来越难以发挥作用。

“我们的辨别思路大致可分为两种，一种是寻找图像编辑痕迹，一般来说这类换脸都会对原先的人脸进行编辑，这个过程中会有类似图片编辑的操作并且留下痕迹，这类痕迹和真正拍摄时的痕迹是不一样的，我们会检测这类痕迹做出记录。另一种则是检测视频中是否会展现出不符合常识的行为，如长时间不眨眼等等。综合上述特征，我们的平台就可以自动判定该内容是否存在深度伪造的可能性，并根据可能性大小来做出相应处置。”萧子豪强调，目前，已经有不少网站开始依据相关规定对 AI 生成的内容进行标识。此举对快速发展的虚拟数字人产业会起到推动、引领作用，并不会影响它的商业价值。

任何一项技术都是一把双刃剑，既能推动社会进步，也能挖掘出人性中隐藏的恶。除了眼下大家有点谈虎色变的“AI 诈骗”，AIGC 的应用非常广泛，例如影视制作、广告营销、电子商务、社交娱乐等。它给社会带来的积极影响也是显而易见的，如提高人们的内容创作效率、孕育新的技术形态与价值模式、产生新的就业机会等等。

任何技术都应该要在可控、可监督的范围内发展无疑是全社会的共识。