转换过程。所以逐渐地，神经网络的设置层越来越多，决策过程成为了一个“黑箱”问题，内部流程非常不透明，以至于该网络的开发者可能都无法完全掌握。目前，最先进的神经网络也缺乏有效的机制能让人理解其推理过程。

比方说，神经网络非常擅长图像识别，当向它们提供足够的数据后，他们可以挑出人眼看不见的图案或差异。利用这一点，深度学习可以实现自动驾驶汽车的行人侦查或肿瘤筛查。但是，当出现超出其参数范围的输入时，神经网络就会出错。静态的、波浪状的人字纹，以及五颜六色的条纹，可能被 AI 自信地识别为“蜈蚣”或“熊猫”。一些常见的图像也会让深度学习人工智能崩溃。把消防车图片倒过来，AI 就会看到一个大雪橇；放大一辆公共汽车的窗户，它在 AI 眼中就变成了一个出气筒。

更可怕的后果还在于 AI 也会存在性别歧视和种族歧视。1982 年伦敦圣乔治医学院的一个博士为了招生的时候筛选学生的效率和公平性，设计了一个程序，他发现机器筛选的结果基本上和人工差不多。管理层觉得这是一个好事，说明人工智能的算法能够代替人。但是很快委员们就发现这里面存在着明显的种族歧视，筛选过程当中如果他们的名字不是白人姓氏，筛选的流程就会不利于他们，而实际上光是没有一个欧洲人的名字就会自动扣除申请者 15 分。

到了 21 世纪，这样的歧视仍然存在。2018 年年初，纽约时报发表文章称，在热门的人脸识别领域，针对不同种族的准确率差异巨大。其中，针对黑人女性的错误率高达 21%–35%，而针对白人男性的错误率则低于 1%。究其原因，主要有两点，一是深色人种数据集的缺乏，二是深色人种人脸特征较难提取。

类似的问题还有很多，包括性别歧视、年龄歧视、种族歧视、地域歧视等等。通过进一步的研究，人们发现这个程序实际上是模仿了人类评估员的行为，最终这个算法并不是自动把种族歧视写进程序里，仅仅是放大了或者模仿了人类评估员的行为。

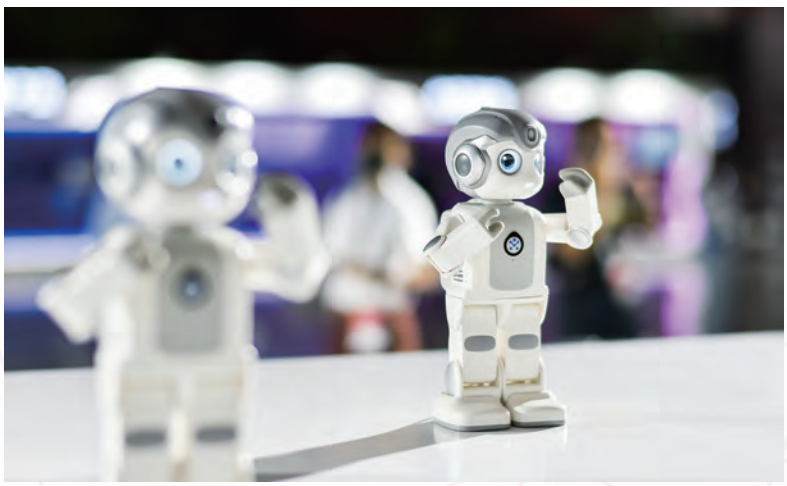
约翰·霍普克罗夫特说，AI 涉及对某一计算机程序进行训练，使其能基于一组训练数据取得良好的表现，然后再将该程序应用到另一组新数据上，使程序能够适应新数据组的过程被称为“泛化”。就是在“泛化”的过程中，AI 程序的偏见出现了。“如果当前大多高层职位都由男性担任的话，那么 AI 程序在执行高层职位人员选聘流程时往往会选择男性。同样，由于亚洲文化与美国文化存在差异，因此，根据中国背景训练的系统可能无法针对美国的问题给出合适的解决方案，反之亦然。”

技术、法律遇挑战

复旦大学计算机学院院长姜育刚一直在从事 AI 的前沿研发工作，他指出，就图文识别技术而言，会出现非常多的技术挑战。“一张很简单的熊猫图片，我们在里面加入一些非常少量的干扰，最后人为视觉看上去还是一个熊猫，但是机器模型就会识别错。还有一些枪的图片，如果加入一些对抗干扰进去，识别结果就会产生错误。比如说自动驾驶领域，如果是限速标牌 80 码，加入一些干扰，被机器识别成 Stop，显然交通上就会引起很大的安全隐患。”

“不只是在图片和视频领域，在语音识别领域，我们任意在语音上加入非常微小的干扰，语音识别系统也可能会把这段语音识别错，这都是可以做到的。在文本识别领域，我们改变一个字母就可以使得文本内容被错误分类，有很多这样的例子。”

当然，这样改头换面、欺骗 AI 的手法还都是一些小伎俩。更高阶的一种叫后门攻击。“我们在训练的时候某一类插入一个后门，比方说用眼镜作为一个后门，用一些技巧训练



右图：人工智能已经进入生活的方方面面。摄影 / 陈梦泽