人工智能背后暗藏多重安全风险

试想一下，有人将一张“贴纸”贴在面部，就可以使人脸识别门禁系统误认为是你，从而轻而易举打开大门；同样是这张“贴纸”，把它贴在眼镜上，就可以1秒解锁你的手机人脸识别，探取隐私如入无人之境。这并非科幻大片的想象，而是首届人工智能安全大赛颁奖典礼现场展示的真实攻防场景。

人脸识别。资料图片

人工智能安全风险已是眼前威胁

前不久，由国家工业信息安全发展研究中心、清华大学人工智能研究院和北京瑞莱智慧科技有限公司等单位联合主办的首届人工智能安全大赛落幕。大赛期间，有关人工智能安全风险引发讨论。与会专家表示，人工智能安全风险已非未来挑战，而是眼前威胁，要重视人工智能安全体系建设，加快促进人工智能安全领域关键技术研究与攻防实践。

人工智能和其他通用技术一样，在高歌猛进的同时，也带来了一定的风险和隐患。曾获“吴文俊人工智能优秀青年奖”的瑞莱智慧首席执行官田天认为，人工智能技术风险发生的范围，正随着应用场景的日趋广泛而逐步扩大，风险发生的可能性也随着其应用频次的增长而持续提高。在他看来，人工智能当前的安全风险主要可以从“人”与“系统”这两个视角来剖析。

从人的视角来评估人工智能的安全问题，首当其冲就是技术的两面性问题，存在人工智能滥用的问题。具体到人工智能的应用中来看，最为典型的代表就是深度伪造技术，它的负向应用风险持续加剧且已产生实质危害。

此次大赛的人脸识别破解演示，所揭示的正是系统的风险，它来自深度学习算法本身的脆弱性。以深度学习算法为核心的第二代人工智能是个“黑盒子”，具有不可解释性，意味着系统存在结构性的漏洞，可能受到不可预知的风险，典型的就比如现场演示的“神奇贴纸”，其实就是“对抗样本攻击”，通过在输入数据中添加扰动，使得系统作出错误判断。

这一漏洞在自动驾驶感知系统同样存在。正常情况下，在识别到路障、指示牌、行人等目标后，自动驾驶车辆就会立即停车，但在目标物体上添加干扰图案后，车辆的感知系统可能会出错，径直撞上去。

大赛期间，《人工智能算力基础设施安全发展白皮书》发布。其中提到，人工智能算力基础设施不同于传统的算力基础设施，推动人工智能算力基础设施安全发展应从强化自身安全、保障运行安全、助力安全合规三个方面发力。

统筹发展和安全，似乎是每项新技术发展过程中面临的必然问题，如何实现高水平发展和高水平安全的良性互动，也是当前人工智能产业发展最为重要的命题之一，现场多位专家就此话题展开讨论。

中国科学院信息安全国家重点实验室副主任陈恺表示，不同于传统的模型修复工作需要重新训练模型，或者依赖于较大量的数据样本，这种方式类似于“微创手术”，只需极少量数据样本，能够大幅提升模型修复效果。

人工智能发展需产学研用多方协同推进

开放环境下的人工智能系统面临诸多安全挑战，如何解决通用人工智能算法全周期的安全保障问题成为重中之重。

北京航空航天大学软件开发环境国家重点实验室副主任刘祥龙表示，从技术上来看应形成从安全性测试到安全性分析与安全性加固的完整技术手段，最终形成标准化的测试流程。

他同时指出，未来的人工智能安全应该围绕从数据、算法到系统各个层次上的全面评测，同时配合一套从硬件到软件的安全可信计算环境。

工商银行金融研究院安全攻防实验室主管专家苏建明表示，人工智能安全治理需要广泛协作和开放创新，需加强政府、学术机构、企业等产业各参与方的互动合作，建立积极的生态规则。政策层面加快人工智能的立法进程，加强对人工智能服务水平、技术支撑能力等专项监督考核力度。学术层面，加大对人工智能安全研究的激励投入，通过产学研合作模式加快科研成果的转化与落地。企业层面，逐步推动人工智能技术由场景拓展向安全可信发展转变，通过参与标准制定，推出产品服务，持续探索人工智能安全实践及解决方案。

事实上，构建人工智能的安全生态，一方面需要技术的持续演进，一方面也需要专项技术人才的建设与培养。田天表示，由于人工智能安全研究目前仍属于新兴领域，专项人才较少，缺乏系统性的研究队伍，此次大赛通过实战演练的方式，验证和提升选手实战能力，为培育一批高水平、高层次的人工智能安全新型人才团队提供了“快速通道”。

专家们认为，从长远看，人工智能的安全问题，还需从算法模型的原理上突破，唯有持续加强基础研究，才能破解核心科学问题，同时他们强调，人工智能的未来发展需确保对整个社会、国家发展的有效性和正向促进性，需要政产学研用多方协同共进。 据《中国青年报》