AI安全攻防战

本报记者 蒋牧云 何莎莎
乌镇 北京报道

AI在掀起科技浪潮的同时，也带来了网络安全、数据安全的挑战。

《中国经营报》记者注意到，在2024年世界互联网大会乌镇峰会（以下简称“乌镇峰会”）上，AI安全挑战成为各方关注的焦点之一，比如会上设置了网络安全技术发展与国际合作论坛、国内首个AI大模型攻防赛等。

多位业内人士向记者表示，随着AI技术不断深入场景，安全事件也愈演愈烈。大模型训练数据泄露、训练遭“投毒”、AI换脸诈骗等问题层出不穷。针对这些现象，不少企业选择通过AI技术解决AI安全问题，通过安全大模型、AI自动化检测、深度鉴伪技术等，提升风险检测反应能力、确保安全防线的牢固。

安全攻防新形势

随着AI技术不断深入场景，安全事件也愈演愈烈。

在乌镇峰会的开幕式上，中共中央政治局常委、国务院副总理丁薛祥发表主旨讲话强调，当前互联网、大数据、云计算、人工智能、区块链等技术不断取得突破，正在全面赋能经济社会发展，但数字鸿沟仍在扩大，网络安全形势更加严峻。

那么，AI带来的安全挑战具体有哪些？全国政协委员、全国工商联副主席、奇安信集团董事长齐向东向记者表示，AI正在黑客攻击、社会操纵和战略规划等关键领域飞速取得进展，并带来前所未有的挑战，安全已经来到临界点。

在齐向东看来，AI带来的安全危机可以总结为“三化”，即黑箱化、黑产化、武器化。具体而言，AI黑箱化将导致内容生成良莠不齐。“在生成AI大模型的过程中，数据和模型在黑箱内部，现实世界和数字世界之间有一道天然隔阂，这不仅使得我们难以洞悉具体使用了哪些数据集和算法，也模糊了攻击者可能采取的具体破坏手段，从而导致有害内容及错误信息的泛滥。”

齐向东谈道：“在今年的实网攻防演练中，奇安信攻击队就成功攻破了某AI大模型，并总结出针对大模型的攻击途径。通过多种手法，篡改了大模型输出内容、让模型出现预料之外或者有害结果，甚至直接瘫痪了大模型。”

AI黑产化则会导致深度伪造泛滥成灾。齐向东表示，不法分子借助AI技术对图像、音视频等内容进行深度伪造，以达到不可告人的目的。无论是公众还是企业，都逃不脱深伪诈骗的陷阱。而AI武器化将导致黑客攻击愈演愈烈。人工智能可以生成恶意软件、钓鱼邮件，也可以快速发现目标系统中的漏洞，大幅降低网络攻击门槛，让不懂代码、不懂技术的普通人也能成为黑客，攻击数量大幅增加。目标处于无法应对的饱和状态，网络空间“易攻难守”成常态。

乌镇峰会期间，由中国图象图形学学会、蚂蚁集团、云安全联盟（CSA）大中华区联合主办的国内首个AI大模型攻防赛亦聚焦于深度伪造的安全问题。蚂蚁集团相关负责人告诉记者：“只需10秒，大模型就能克隆声音、复刻照片，甚至能生成‘你’的视频，从而引发深伪欺诈、色情影像伪造、假新闻等社会事件；大模型‘越狱’问题频发，诱骗AI听从不怀好意的指令，生成血腥、暴力、歧视、仇恨的图片、视频，危害网络空间安全。”

安恒信息相关负责人也告诉记者：“随着AI技术不断深入场景，安全事件也愈演愈烈。大模型训练数据泄露、训练遭‘投毒’、AI换脸诈骗等问题层出不穷。从训练到应用，大模型的安全风险无处不在。比如数据隐私和安全问题，AI系统通常需要大量的数据来训练模型，这可能涉及敏感的个人信息。如果这些数据被不当使用或泄露，将对个人隐私造成威胁。还有内容安全问题，如何确保AI不生成错误内容、违规内容和恶意代码？如何确保大模型生成的代码或内容不被用于执行安全攻击？这都是AI带来的新的安全挑战。”

以AI治理AI

“以AI治理AI”模式已经成为解决安全风险的一大趋势。

中国工程院院士、中国图象图形学学会理事长王耀南指出：“加强大模型安全保护，构建完善的安全防护体系，是确保人工智能技术持续、稳定、健康发展的关键所在，也是我们在这个充满机遇与挑战的时代必须肩负起的重要使命。”

记者在乌镇峰会会场了解到，针对AI技术带来的深度伪造风险，蚂蚁集团正通过蚁天鉴和ZOLOZ等安全技术产品加强对图像、视频的鉴真能力。据介绍，蚁天鉴不仅支持图像、视频等多模态内容真实性及深度伪造的监测，还支持大模型X光、大模型基础设施测评、应用安全测评、围栏防御等技术能力。记者现场从图库中选择了数张照片，仅用几秒，蚁天鉴就可以准确识别图片或视频片段的真伪。

而ZOLOZ则更专注于攻克AI换脸难题，其人脸识别准确率达99.9%。在现场，观众可以上传一张个人照片，由AI基于照片合成新的人脸图像来试图突破ZOLOZ防御系统。工作人员告诉记者，目前ZOLOZ已为中国、印度尼西亚、马来西亚、菲律宾等24个国家和地区提供技术服务，涵盖金融、保险、证券、信贷、电信、公共服务等多个领域，累计服务用户超12亿。

齐向东告诉记者，未来强化人工智能安全治理，需要重点采用三大技术策略。第一个策略是结合大模型基础运行环境、训练环境、API接口以及数据安全进行多维度、体系化防护。大模型生命周期的每个环节都存在大量不确定性，无论是数据安全、算法开发和模型安全、内容还是应用安全等方面，都要做到合规。

“以AI治理AI”不仅是技术上的需要，也是在确保AI系统安全性和可靠性方面的重要战略。 刘洋/制图

第二个策略是用“鉴伪”“防伪”技术有效遏制深度伪造。针对正在野蛮生长的生成式伪造语音技术、生成式伪造视频技术，应该尽快发展相关检测技术。比如，奇安信自研的深度鉴伪模型能够准确识别多种前沿AI伪造技术生成的虚假图片视频；洛基平台可以通过内网在线访问，上传图片、视频开展深度鉴伪。

第三个策略是用安全大模型反哺安全能力大提升。建立体系化的安全防护系统，是AI安全大模型驱动安全的重要前提。奇安信的内生安全体系，把网络安全设备和业务流转、不同层次的信息系统有机结合起来，做到安全能力的无死角，确保多道网络安全防线有效协同，实现从宏观管控到微观检测的全面防护。记者了解到，奇安信在内生安全体系之上，部署了自研的QAX-GPT安全大模型，这样不仅让大模型更懂客户业务，同时也让安全体系效率更高、能力更强。经过反复训练打磨，AI安全大模型的研判效率已经提到了人工的60多倍。

事实上，类似的“以AI治理AI”模式已经成为解决安全风险的一大趋势。安恒信息相关负责人告诉记者，公司正从对抗性训练、自动化检测、大模型风险检测、联邦学习和隐私保护、AI辅助威胁情报等5大方向进行探索。其中，对抗性训练通过在模型训练过程中引入对抗性样本，目前“恒脑”利用该技术来微调训练模型，使“恒脑”能够更好地抵御对抗性攻击。这种方法提高了AI系统对恶意输入的鲁棒性。

大模型风险检测方面，安恒则通过恒脑智鉴针对大模型风险评估采用精细化风险评估方法，覆盖12大内容安全风险领域，细分为40余种小类，确保无遗漏。同时，配备20余种检测手段和超过25000个测试用例，提供详尽的数据分析和安全报告，能够帮助政企监管机构快速、精准地发现潜在问题并采取相应措施。“这些探索和实践表明，‘以AI治理AI’不仅是技术上的需要，也是在确保AI系统安全性和可靠性方面的重要战略。随着AI技术的不断发展，这一领域将继续演进和扩展。”该负责人表示。

未来，AI攻防还有哪些深化的方向？安恒信息相关负责人告诉记者，过往业内比较关注数据投毒，但安恒信息研究发现，相比数据投毒，大模型权重文件投毒后门的适用性更广，危害性更大。通过模型权重文件投毒方式，模型可被控制会遵从恶意控制者行动，平时是一个常规大模型，当恶意控制者在任意时候发送指令，即可马上让它变成恶意大模型，如何在任何时候让模型可控、不越界是安恒接下来要研究和探索的方向。

互联互通十年再起航 更多重磅政策陆续落地

本报记者 郭婧婷 北京报道

“2012年，沪港两所高层共聚深圳的一个茶馆，在一张小小的餐巾纸上画出了沪港两地股票市场交易联通的雏形路径，这正是沪港通的最初蓝图。两年后，沪港通成功推出。”11月18日，上海证券交易所总经理蔡建春在“互联互通十周年高峰论坛”上讲述机制设立之初的故事。

根据港交所最新数据，截至2024年9月底，2014年至2024年沪深港股通总成交额为177万亿元；内地投资者通过港股通持仓总值为3.3万亿港元，较2014年年底高出200多倍。沪深港通下合资格股票超过3300只，已覆盖沪深港三地市场上市公司总市值的九成、成交规模的八成以上。

中国证监会副主席李明11月18日在互联互通十周年高峰论坛上表示，中国证监会将继续优化完善互联互通机制，拓展内地企业境外上市渠道，扩大期货市场开放，加快推动新一轮全面深化资本市场改革开放，深化机构产品合作，加强两地监管合作，共同维护两地市场健康稳定发展。

让境内外资本更畅通

梳理互联互通近年来的发展轨迹：2021年10月18日，港交所推出MSCI中国A50互联互通指数期货；2022年7月4日，互联互通投资标的启动交易两地交易所符合条件的ETF纳入；2023年3月13日，沪深港通合资格股票范围再次扩容，交易标的首次纳入在香港主要上市的合资格外国公司，并新增超过1000家A股上市公司；2023年5月15日，北向互换通正式上线，互联互通拓展至新的资产类别，首次实现衍生品跨境互联互通，为国际投资者提供了精准高效的人民币利率风险管理工具；2024年4月19日，中国证监会公布五项资本市场对港合作措施，包括放宽合资格ETF范围、将REITs纳入沪深港通，以及将人民币股票柜台纳入港股通等；7月9日，中国人民银行宣布支持境外投资者使用债券通持仓中的在岸国债和政策性金融债作为北向互换通的履约抵押品。

由此，互联互通机制在不断完善发展中迎来首个十年成绩单。

11月11日，港交所发布《内地与香港资本市场互联互通十周年白皮书》（以下简称《白皮书》）显示，过去十年，沪深港通成交活跃度稳步提升。2024年前三季度，北向和南向交易的日均成交额分别为1233亿元和383亿港元，与2014年开通首月相比，分别增长21倍和40倍，已占到内地市场成交总额的6.7%和香港市场成交总额的16.9%。

据港交所行政总裁陈翊庭介绍，目前沪深港通已成为国际投资者交易和持有内地A股的主要渠道，有接近77%的外资通过这个渠道持有内地股票。此外，在国际投资者投资内地债券市场的交易中，一半以上通过债券通进行，债券通北向通也成为国际资本投资内地债券市场的主渠道。

港交所数据显示，自2022年7月互联互通下的ETF交易启动以来，合资格产品范围稳步扩大，覆盖的指数也更加丰富。北向合资格ETF已由启动之初的83只增长至225只，成交额占沪深市场ETF总成交的30%以上；南向合资格ETF已从4只增加至16只，成交额占香港市场ETF总成交的97%，覆盖了宽基、行业主题、策略等不同类型指数。

那么，互联互通，“通”的是什么？如何让境内外资本更为畅通？对此，南开大学金融发展研究院院长田利辉接受《中国经营报》记者采访时表示，要强化桥梁作用，推动机制扩容和更好推进人民币国际化。要促进香港作为连接内地与国际市场的桥梁作用，实现更广泛的互利共赢。也要使香港在国家金融市场开放进程中扮演更有效的角色，如“防火墙”“试验田”及资金引领平台。还要充分利用互联互通机制的优势，稳健推进人民币国际化进程。

官方数据显示，自2014年沪港通启动以来，北向交易累计为A股市场带来近1.8万亿元资金净流入。从历史数据看，在开通以来的2200多个沪股通和深股通交易日中，北向资金在约45%的沪深300指数下跌交易日逆势净买入A股，发挥了稳定预期及对冲风险的积极效果，对波动中的A股市场起到了一定支撑作用。

互联互通交出的成绩单，同两地证监会和相关部委的支持分不开。过去十年来，有关部门参考国际成熟市场的经验，结合沪深港三地的实际情况，出台了一系列规范性文件。在税收政策方面，财政部、国家税务总局等部门也多次发布政策，对互联互通相关税收安排和优惠措施予以明确，有效降低了跨境投资成本，在鼓励投资者积极参与互联互通、提振市场信心和活跃度方面发挥了重要作用。

进一步丰富产品类别

作为互联互通资本市场参与者，金融机构从业者感触颇深。

“对国际资本而言，投资的第一个挑战是‘可及性’。投资之前，对于这个市场是否向我们开放、是否需要得到一些许可，或者得到了多少配额、有什么限制等，是许多资管机构都面临的疑问，互联互通正好解决了这些问题。”互联互通高峰论坛圆桌环节中，泰康资产管理（香港）有限公司总经理张乐如是表示。

“作为全球第二大经济体，中国金融市场的外资参与度仍然比较低，目前外资持有内地股票和债券的比例都不到5%，还有很大的提升空间。”陈翊庭坦言。

受访业内人士表示，对于金融机构来说，互联互通带来了巨大的机遇和挑战。首先，互联互通机制提供了更多的投资机会，可以更好地把握国内外市场的投资机会。其次，互联互通促进了金融机构的业务拓展和创新，机构需要不断适应市场变化，提高自身的专业能力和服务水平。然而，互联互通也带来了一定的风险和挑战，如市场波动、汇率风险、合规风险等，机构需要加强风险管理，确保投资和交易的合规性和稳健性。

《白皮书》显示，截至2024年9月，十年间共有260多家香港经纪商参与沪深股通业务，开展港股通业务的内地证券公司超100家，沪深港通业务已成为两地证券经营机构重要的业务组成部分。同时，为满足内地投资者的跨境投资需求，追踪港股上市公司的研究报告覆盖面和发布密度均有显著扩大和提升。

招商基金资深策略分析师黄亮观察到，从2015年起，内地的公募基金、保险资金、年金基金等都陆续获得了港股通业务资质，中长期资金开展境外资产配置成为常态化，带动了内地投资者参与香港证券市场的积极性。目前，通过港股通持有的证券资产总市值超过3.3万亿港元。互联互通也让国际投资者能进一步投资中国资本市场，进而对我国A股和国债纳入MSCI、富时罗素等全球主流指数起到了推动作用，这些指数也让更多境外投资者增配优质中国资产。

Wind数据显示，截至10月，今年已有170多家企业完成境外上市备案，远超过去年全年的72家，赴港上市成为众多企业的选择。

田利辉表示，香港作为与内地关系最紧密的国际金融中心，其法律、市场规则及投资者结构与内地更具兼容性，备案流程更易标准化。相较之下，赴美或其他市场上市需考量更复杂的法律和监管环境。差异化管理既有助于提升备案效率，又可防范境外市场法律与监管风险，保障企业合法权益与资本市场稳定发展。

展望未来，港交所方面表示，将进一步丰富产品类别、扩大标的范围，持续优化沪深港通交易机制和配套服务，探索更多有助于提升投资者参与度和便利性的措施，不断完善债券通和互换通等安排。

记者注意到，随着境内外投资者的沪深港通参与程度不断加深，不少投资者希望可以在沪深港通下参与大宗交易，提高成交效率，降低市场价格波动对交易的影响。2023年8月11日，两地证监会宣布就推动大宗交易纳入互联互通机制达成共识。

“目前，两地交易所和结算公司正在积极开展REITs纳入沪深港通的技术和市场准备工作。未来，可以进一步探索将合资格ETF标的的底层资产从股票拓展至更多资产类别等。”港交所透露前述五项资本市场对港合作措施的实施进展。