2023年11月24日 星期五
责任编辑:严粒粒 版式:戚建卫 联系电话:0571-85311516 邮箱:zjrb@8531.cn

一项安全科技领域研究成果,获浙江省科学技术进步一等奖——

用AI之“道”,降伏AI之“魔”

本报记者 谢 晔 张 留

人工智能带来的颠覆性安全挑战正在加快凝聚人们的共识。

11月8日,国家主席习近平向2023年世界互联网大会乌镇峰会开幕式发表视频致辞指出,“我们倡导安危与共,构建更加和平安全的网络空间”“深化网络安全务实合作,有力打击网络违法犯罪行为,加强数据安全和个人信息保护”,在全球引发热烈反响。

两天后,2022年度浙江省科学技术奖揭晓,由蚂蚁集团联合清华大学攻克的安全科技领域研究成果《面向海量交易的分布式协同风险防御关键技术与系统》(以下简称分布式风险防御技术),获得了浙江省科学技术进步一等奖。

作为多年稳居“中国科学院区域创新能力报告”前10的科创大省,浙江的“一等奖”历来是国家科技大奖的重要候选。今年浙江这个一等奖,又因切中时代痛点,而更加备受瞩目。它究竟练成了什么“独门秘技”扛住了愈加严峻的安全风险与挑战?又预示着平台经济乃至浙江省在国际最前沿科技探索活动中,具备多强的实力、怎样的担当?

日前,记者采访了该项目的主要专家团队:蚂蚁集团副总裁、大安全事业群总裁赵闻飙,清华大学计算机系副主任徐恪,蚂蚁集团研究员、大安全事业群首席科学家王维强。他们的回答是:用AI之“道”,对抗AI之“魔”。

大模型时代,给安全防御带来巨大挑战。(蚂蚁集团供图)

AI对抗AI
练成“金刚不坏”之身

先来看看由多位院士组成的鉴定委员会是如何评价的——

总体评价:该项目系统复杂度高、问题难度大,技术创新性强。

对“一个系统”和三项核心技术的评价:其端边云分布式协同风险防御体系为国际首创,基于对抗学习的人工智能模型抗噪防御技术、可扩展的在线动态图表征对比学习技术、知识驱动的自适应多维主动询问风险分析技术达到国际领先水平。

“院士们的鉴定有两层含义。”徐恪向记者解释:

其一是“国际首创”,指的是端边云分布式协同技术,比如中国的移动支付技术和需求全球领先,需要协同移动端、云计算等技术,还要保障各方的隐私数据安全,这种协同端、边、云的分布式协同风险防御体系,在国际上此前还没有出现过。

其二是“国际领先”,指的是针对人工智能驱动的对海量数据的攻击,人工防御已经不再现实,只有通过采用人工智能抗噪防御、在线动态图表征对比学习等手段,对风险漏洞进行智能化的自动检测,堵住随生随灭的漏洞。

在大模型时代,由人工智能驱动的“魔”,更是对传统防御技术形成了“降维打击”,不仅可能引发技术崩溃,更带来一系列社会、伦理、道德、法律等全新挑战。“如果不加以有效应对,互联网的风险敞口会像决堤一样。”赵闻飙说。

首先,是内生安全问题。如果给数据库不断投喂带有特定价值观的数据,那么很容易就产生“数据偏见”“观点霸权”等问题。前不久,某公司推出的人工智能学习机被发现一篇课文《蔺相如》的读后感中,含有歪曲历史等违背主流价值观的内容。当事企业事后查明,内容是第三方引入的,也就是无意或有意向数据库中引入了有害数据,输出“含毒”结果。

其次,是衍生安全问题。例如,被恶意攻击者借助人工智能技术,可以生成假照片、假护照等实现AI伪造,还可以模仿人脸、声音等进行诈骗,甚至逼真模仿身份骗过安全管理员,取得顶级密钥。“人工智能技术脆弱性被恶意利用,就会让安全问题变得更加棘手。”赵闻飙说。这样的事件时有发生。今年,就有杭州市民借助ChatGPT杜撰交通取消限行的新闻,闹得满城风雨。此外,据业内不完全统计,仅2022年,我国AI诈骗案件就达到了50万件,涉及金额超过100亿元。

用AI对抗AI来实现“邪不压正”,首要任务是抗得住来自四面八方仿真度极高的智能攻击,练就“金刚不坏”之身。专家们从《射雕英雄传》的周伯通“左右互搏”功夫中受到启发。

在蚂蚁集团的技术中,也实践了“左右手互搏术”:一边模拟攻击、一边自卫防守。“我们在开发和迭代分布式风险防御技术时,每天都会模拟各种各样的攻击。”王维强说,假设一个真实的人,一天能模拟200次针对AI模型的攻击,那么蚂蚁集团现有的技术力量已经可以实现2500个人每天模拟发起50万次攻击。

在海量的一攻一防之中,蚂蚁集团练就了风险防御的“绝顶功夫”,支撑风险防御技术与系统历经了六次迭代,最终升级为“可信AI+大数据驱动”的智能安全防御系统,在网络安全这条赛道上,跑到了全球领先的位置。

厚积薄发
打造数实融合安全基石

几个月前,徐恪在首届国际基础科学大会上,有了一个新发现:这个大会把理论计算机与信息科学也列入了基础科学的范畴。“的确,从工程到制药,千行百业都在利用大数据、云计算提升生产力。”

基于中国数字经济的领先优势、制造业雄厚的实力,“数实融合”成为当前数字经济发展的“必选项”,而新型的网络安全风险给数实融合带来了前所未有的挑战。蚂蚁集团和清华大学联手取得这个“一等奖”的巨大价值就此凸显。

“以前AI只是一个工具,现在AI正在引领加强原创性、引领性科技攻关,它不仅仅只是一项新技术,更是开启一个新世界的密钥,最终将会嵌入整个经济运行体系中。我们开发的分布式风险防御技术就是为了打造一座‘数实融合’的安全基石。”赵闻飙说。

“安全科技是数字技术的重要组成,同样要做到顶天又立地。”在赵闻飙看来,就像一个硬币的两面,安全科技的突破和经济效益的提升,就是矛盾统一的,安全不仅仅是防御,更能促进发展和创新。加速基础科学研究以寻求颠覆性突破,探索人工智能与安全技术深度融合,是应对当前安全挑战的华山一条路。

新型网络安全风险已经存在于数据要素流转的所有环节中。蚂蚁集团从最初因为海量交易数据的安全而涉足安全领域,一步步从专家经验,到浅层机器学习、深度学习、预训练模型,再走到今天的基础大模型的演变和应用,历经六次迭代,步入了“国际首创”的“无人区”。自身也从企业需求出发,自觉扛起了引领性科技创新的国家战略担当。蚂蚁集团已与全球30多所知名院校科研合作,并组建了九大安全实验室,开展前沿安全科技研究。

在数据流传的各个关键节点,蚂蚁集团安全科技都有出色的独创性成果。比如,能够自动挖掘关键数字化基础设施及供应链中潜在未知漏洞的“天穹实验室”,获得了国家网络安全优秀创新成果大赛、天府杯国际网络安全大赛等多个奖项;用独特的隐私算法实现数据“可用不可见、可算不可识”的“隐语”系统,已经广泛运用到了金融、医疗等领域,有效防范金融诈骗事件发生。

“数字时代,技术变革太快了,没有一招鲜的解决方案了。”王维强说,尤其是大模型面世后,在训练过程中需要安全监控,在流入市场时需要质检,在进入市场后仍然需要可控的使用方式。

针对大模型时代的机遇与挑战,蚂蚁集团正通过大模型的涌现能力探索下一代安全科技:一是用大模型保障安全,比如基于安全基座模型能力,持续升级全方位全链路的智能风险防御体系;二是保障大模型技术本身的安全,蚂蚁集团自主研发了大模型安全一体化方案“蚁天鉴”,同时也在今年世界人工智能大会上获得了“镇馆之宝”奖项。

所谓“蚁鉴”,就是大模型安全测评方案,通过诱导式对抗生成技术,对大模型进行诱导式问答,找出弱点和漏洞。所谓“天鉴”,就是大模型风险防御方案,基于智能风控技术,帮助大模型挡住恶意提问。

“一旦大模型在细分领域大规模应用,那么‘蚁天鉴’就有可能成为另一块基石。”赵闻飙认为,安全科技研究需要长期坚持、久久为功。面对生成式人工智能正在带来的深度智能时代,安全科技需要随着人工智能的对抗升级而不断迭代、不断完善,蚂蚁集团与清华大学等专家学者要做的还远未穷尽。

向实而生
面向生态抓机遇

刚刚过去的双十一,一些中小企业也许会有这样的困扰:除了在网络平台投放的营销成本被“薅羊毛”,还有流量造假、数据欺诈等等诸多问题。

“我国有几千万中小微商家和企业,都需要提升数字化经营的安全能力。”王维强说,蚂蚁集团一直在坚持向实而生,最典型的是通过免费服务的方式,将诸多安全能力给到生态小微企业,帮助他们解决数字化转型中的“不敢”和“不会”。

在王维强看来,数字科技企业向实而生,有助于增强实体经济的韧性和活力,也有助于充分激活数字技术的价值。因为单打独斗已经不适应这一时代的科技创新和产业变革,安全更需要面向应用和生态,高效技术和产业协作是抓住技术变革机遇的关键。

近一两年,不少手机厂商与蚂蚁集团合作研发了一项名叫“可信隐私沙盒”的技术,这个沙盒(记者注:计算机专业术语,一种安全机制)部署了用于支付宝科技反诈的技术能力。沙盒当中只要部署“电信诈骗可疑感知模型”,就能使得风险感知、电诈定性均可在用户个人的手机上完成,从而实现更有效的源头上风险治理。

这种向实而生的拓展性,是可以预见的无限广阔。单单这款“可信隐私沙盒”,通过“科技创新+产业联防”的方式,可以通过手机系统进一步赋能各类应用,提高整体风控水平,未来还会运用到更多场景、更多产业中去。“向实而生的数字技术,确实正在成为实体经济的基础设施。”徐恪介绍,他曾创作《赛博新经济》一书,重点阐述了互联网产生以来对经济系统带来的重要影响。由蚂蚁集团所代表的数字经济中坚力量,已经成为经济运行体系中的基础性支撑,在许多领域向世界输出了合作范例。

事实上,蚂蚁集团就在发挥这样的作用。

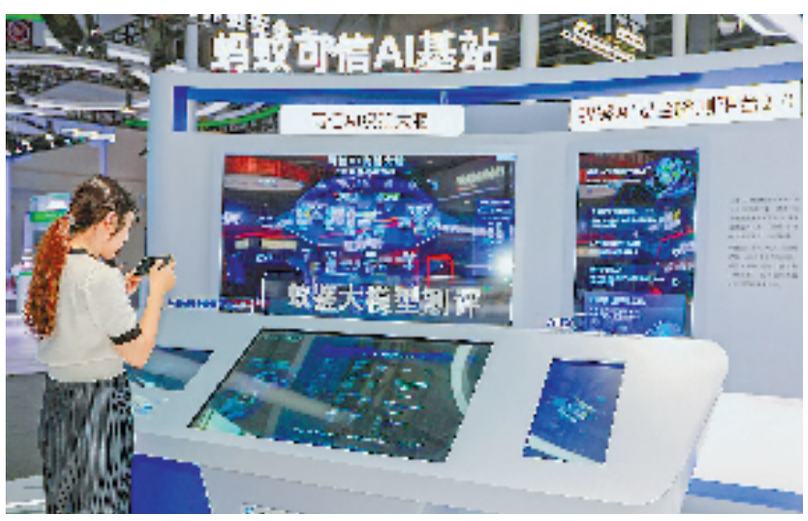
横版、竖版、圆形、三角形,工作证、学生证……在拥有30多种官方证件的菲律宾,如何让市民用上安全的电子钱包?

“针对菲律宾证件繁多的情况,我们设计了一套通用开发框架,来适配不同版式的证件,把注册时间由原来的最快5天缩短到3分钟左右。”赵闻飙所举的案例,是蚂蚁集团利用技术优势帮助东南亚地区进入移动支付时代的一个缩影。

为了响应共建“一带一路”,蚂蚁集团输出身份安全、业务安全、产业风控、合规安全等全链路的安全风控产品与服务,支撑印尼、马来西亚、菲律宾等14个国家和地区的移动支付安全保障及数字金融普惠。

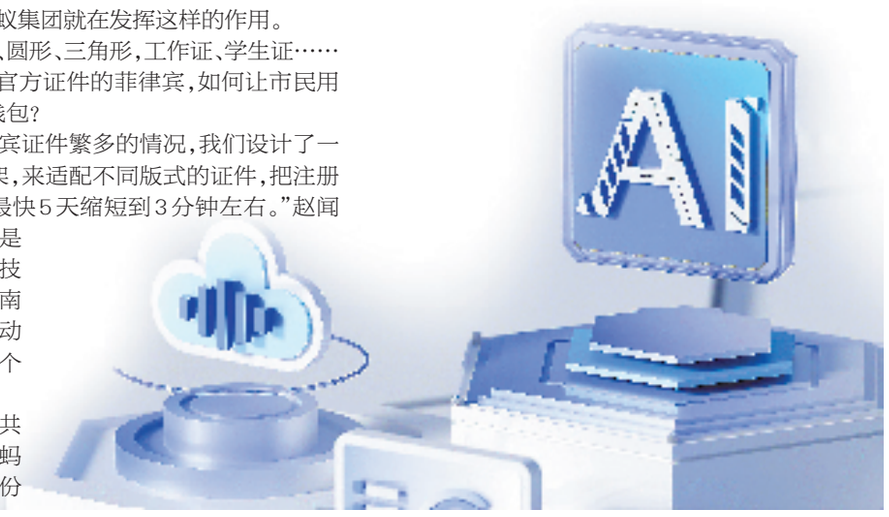
“走出去”背后的技术支撑,则是以人像目标为核心的视觉智能研究。今年,由蚂蚁集团与清华大学共同完成的“无约束人像目标智能感知与理解”研究,获得了“吴文俊人工智能自然科学奖”一等奖。这项研究揭示了人脸目标“不变性特征学习”的重要性等三大重要科学发现,向行业开放的人脸识别平台,在实现把人脸身份识别精度大幅提升的同时,可将误报率降低4倍。目前,蚂蚁集团借此研发的可信身份认证产品与服务,已经成为多国互联网APP的第一道“门禁”,并且覆盖了电信、金融、公众服务等10多个行业及领域。

更可想象的向实而生正在发生。随着东南亚地区移动支付能力的建设,蚂蚁集团在中国与东南亚、东南亚与东南亚之间,打开了一条资源要素流动的通道,这是提升开放能级的关键基础设施,甚至可能将中国的企业、中国的产业带去更广阔的天地。



蚂蚁集团联合清华大学展出蚁鉴安全检测平台。(蚂蚁集团供图)

“天穹实验室”团队。(蚂蚁集团供图)



记者手记

重视安全科技的极端重要性

谢 晔 张 留

随着人类快步进入人工智能时代,网络世界的风险可能不断向经济社会各领域渗透、扩大,甚至影响国家安全。因此,人类社会越是智能,就越有必要重新审视安全科技的极端重要性。

这种极端重要性,从安全科技研究范式的转变中就可感知。从过去重应用,到现在重基础研究,安全科技已经是支撑和保障各领域安全的逻辑起点。只有积累了足够多先进的、底层的技术,才能在最极端事件发生时确保万无一失。安全也不仅仅是防御,更能促发展,成为创新的新动能。

在采访中,我们发现,蚂蚁集团的安全技术起源于支付,但并不局限于支付。这十几年来,在安全科技领域,蚂蚁集团已与全球超30所知名院校、100多名顶尖学者团队进行科研合作,一直把安全作为基础研究在做,持续攻坚业务安全、网络安全、数据安全、内容安全、人工智能安全等很多安全方向,也取得了不少原创成果。正因如此,这次获得浙江省科技进步一等奖的项目,并不是某一项技术,而是一整套的风险防御体系。这离不开企业持续坚定地投入安全科技的基础研究与技术突破。

不过,现在这个时代,风险的变化和技术变革式的应用更新速度太快,很难再有一招鲜的解决方案。因此,未来需要有更多的底层技术突破瓶颈。我们期待,更多企业能够站在技术迭代的最前沿,研发出更多“独门秘籍”,用更高水平的安全科技保障更高质量的发展。

科技速递

全球首个CRISPR基因编辑疗法获批

英国监管机构近日说,已批准基于CRISPR基因编辑技术开发的Casgevy疗法投入应用,用于治疗两种血液病,这是全球首个获批应用的CRISPR基因编辑疗法。

英国药品与保健品管理局在当天发布的一份声明中说,在经过安全性、有效性等方面的严格评估后,该机构已批准品牌名称为Casgevy的基因编辑疗法投入应用,用于治疗12岁及以上的镰状细胞病和输血依赖型β地中海贫血患者。

这种疗法由美国福泰制药公司与瑞士CRISPR治疗公司共同开发。该方法通过从患者的骨髓中提取干细胞,并在实验室中编辑细胞中的基因,然后回输到患者体内,使患者身体能够产生功能性血红蛋白。

镰状细胞病和输血依赖型β地中海贫血都是由血红蛋白基因错误引发的遗传性血液病。镰状细胞病在具有非洲或加勒比族裔背景的人中多见,患者会经历非常严重的疼痛、严重且危及生命的感染,以及贫血。英国约有1.5万人患有镰状细胞病。输血依赖型β地中海贫血主要影响地中海、南亚、东南亚和中东地区的人群,可能导致严重贫血,患者通常每3至5周需要输血一次。

根据英国药品与保健品管理局的声明,临床试验数据显示,镰状细胞病患者接受这种基因编辑疗法后至少12个月没有出现严重疼痛的比例达97%,输血依赖型β地中海贫血患者接受这种基因编辑疗法后至少12个月不需要输入血红细胞的比例为93%。

此前,骨髓移植是这两种疾病患者唯一的永久性治疗选择,但移植的骨髓必须来自匹配的捐赠者,并仍存在排异风险。

CRISPR全名为“成簇的、规律间隔的短回文重复序列”,原本是细菌防御病毒侵入的一种机制,被科学家用于编辑基因。法国科学家埃玛纽埃勒·沙尔庞捷和美国科学家珍妮弗·道德纳因为开发出相关技术而获得2020年诺贝尔化学奖。这项技术已成为可高效、精确、程序化修改细胞基因的工具。 据新华社消息

日本推动设立宇宙战略基金

日本政府内阁会议近日确定了关于日本宇宙航空研究开发机构的法律修正案,计划在这一机构设立宇宙战略基金,以支持企业和大学的航天技术研发。

综合日本媒体20日报道,日本政府力争在本届国会通过此修正案。设立宇宙战略基金,目的是推动民间力量开展航天开发,将向公开招募遴选出来的企业、大学等提供资金支持。

据悉,该基金将在今后10年内向企业和大学等提供1万亿日元(约合67亿美元)的资助资金。

日本今年6月修订的宇宙基本计划提出,要强化日本宇宙航空研究开发机构战略性且灵活的资金支持功能,培育具有竞争力的航天企业。设立宇宙战略基金是迈出的第一步。

日本媒体报道说,日本将新制定“宇宙技术战略”,以卫星、宇宙科学和探测以及火箭等为支柱。日本宇宙航空研究开发机构预计将以这份战略为依据来选定基金资助对象。

据新华社消息

2023年世界无线电通信大会开幕

2023年世界无线电通信大会近日在阿联酋迪拜拉开帷幕。来自国际电信联盟193个成员国、相关国际组织和相关企业的4000余名代表参加会议。

本次大会会期4周,共设立28项议题,涉及5G/6G新增频率划分、北斗短报文服务系统全球应用、卫星互联网未来可持续发展、航空和航海现代化频率使用、气候变化与气象探测频率使用等内容,并研究各主管部门提出拟设立的52项未来大会议题。

会议期间,中国代表团成员将在国际电信联盟规则框架下,积极参与《无线电规则》制修订和大会议题的讨论,广泛开展国际交流,为推动共建无线电领域全球命运共同体,促进国际或区域频谱管理事宜达成共识贡献中国智慧、中国方案、中国力量。

世界无线电通信大会是由国际电信联盟主办、国际无线电事务立法缔约的最高级别会议,每3至4年召开一次。 据新华社消息