大模型落地，跑得快更要跑得稳

AI世界

◎本报记者 崔 爽

比盛夏的上海更火热的是2024世界人工智能大会暨人工智能全球治理高级别会议（以下简称“WAIC 2024”）。大会线下参观人数突破30万人次，创历史新高。

值得注意的是，WAIC 2024的首发首秀不仅涉及模型更新换代，还涵盖应用、平台、系统等。行业和观众的目光更多投向与模型落地紧密相关的交互体验、商业模式等领域。

一个引发广泛关注的问题是，随着大模型能力不断增强，其安全性、可靠性、可控性也日益受到挑战。尤其是面对行业用户合法合规、精准可控等要求，大模型可能存在的数据安全、幻觉等成为绕不过的问题。

中国信息通信研究院华东分院人工智能事业部主任常永波说，应用价值与应用安全是大模型发展的两翼，当前大模型已进入快速迭代期，在积极探索落地应用的同时，大模型厂商也要高度重视应用场景需求下对安全的行业要求。

在WAIC 2024上，参观者在某大模型演示屏前体验交流。 新华社记者 方喆摄

技术自身缺陷不容忽视

依托庞大参数规模、海量训练数据、强大算力资源，大模型作为人工智能领域最热门的技术分支，已在多个领域表现出超越人类的能力。

“金融、医疗、教育、政务、制造等众多领域都在积极探索大模型安全应用范式，以应对大模型安全风险。”常永波介绍，伴随大模型的深度应用，产学研用各方都在加强大模型安全威胁和防御技术体系研究。在原有可信人工智能治理体系框架基础上，提升大模型的鲁棒性、可解释性、公平性、真实性等能力成为行业研究热点。安全评测技术和安全防御技术的不断成熟，有效护航大模型发展。

WAIC 2024上，清华大学、中关村实验室、蚂蚁集团等机构联合撰写的《大模型安全实践（2024）》白皮书（以下简称“白皮书”）正式发布。白皮书显示，大模型技术存在自身缺陷，包括生成内容不可信、能力不可控以及外部安全隐患等问题。

“幻觉是大模型目前比较难解决的问题。”常永波说，模型在遵循语法规则的同时，可能产生虚假或无意义的信息。这一现象源于大模型基于概率推理的输出方式。它可能导致对模糊预测的过度自信，从而编造错误或不存在的事实，影响生成内容的可信度。“智能涌现”是大模型的另一种效应，它可以让模型展现出色性能，也具有突发性、不可预测性和不可控性等特征。

另外，大模型的脆弱性和易受攻击性使外部安全隐患难以消除。相关数据显示，随着大模型技术快速发展，相关网络攻击也在增多。

聚焦安全可靠可控性建设

大模型带来的种种风险，对监管方、学术界、产业界是全新且不可回避的问题。

近年来，《互联网信息服务算法推荐管理规定》《互联网信息服务深度合成管理规定》《生成式人工智能服务管理暂行办法》《科技伦理审查办法（试行）》等政策法规相继发布，搭建起我国人工智能治理的基本框架。一系列政策法规坚持发展与安全并重原则，强化科技伦理风险防控，从技术发展与治理、服务规范、监督检查与法律责任等层面对大模型安全发展提出要求。

白皮书提出，构建大模型安全政府监管、生态培育、企业自律、人才培养、测试验证“五维一体”的治理框架。

在监管方面，常永波介绍，敏捷治理正成为一种新型治理模式。该模式以柔韧、流动、灵活及自适应为特点，倡导多元利益相关者共同参与，能快速响应环境变化。在实施治理策略时，结合柔性伦理规范和硬性法律法规，构建完善的治理机制，在规制大模型风险的同时平衡创新与安全。

“为确保大模型在实际应用中发挥最大效能，防止潜在风险和滥用，大模型建设通常会聚焦三个重要维度：安全性、可靠性和可控性。”蚂蚁集团安全实验室首席科学家王维强解释，安全性意味着确保模型在所有阶段都受到保护，防止任何未经授权的访问、修改或感染，保障人工智能系统无漏洞、免诱导；可靠性要求大模型在各种情境下都能持续提供准确、一致、真实的结果，这对于决策支持系统尤为重要；可控性关乎模型在提供结果和决策时能否让人类了解和介入，以便人类根据需要进行调适和操作。

王维强特别提到时下备受关注的Agent（智能体）。他说，Agent是目前大模型落地的关键路径，但复杂的Agent体系进一步扩大了大模型风险敞口。目前RAG（检索增强生成）、指令遵循、知识图谱嵌入等方法可有针对性地提升模型输出的可控性和准确性。

合力推动人工智能健康发展

“目前来看，让大模型完全不犯错几乎不可能，但减小犯错几率，减弱错误危害性，是可以做到的。”常永波说，安全治理需产学研共同发力，中国信息通信研究院已开展一系列标准和测评研究，头部厂商也在加速构建自身的安全和治理体系。

蚂蚁集团安全内容智能负责人赵智源介绍了相关经验。一方面，在大模型产品投入应用前，企业需做好全面评测，对暴露出的安全问题展开针对性防御，把好入口关；相关产品进入市场后，也要时刻监控可能出现的风险隐患，进行技术补救和改进。另一方面，模型技术通常跑在安全技术前，行业研究要保持一定前瞻性。

“我们很早就开始探索基于安全知识构建视觉领域生成内容风险抑制的技术。在多模态大模型发布后，我们又将这一技术集成到多模态基座中，降低风险内容生成比例。”赵智源介绍，蚂蚁集团已构建起面向产业级应用的大模型安全一体化解决方案“蚁天鉴”2.0版本，形成包括大模型基础设施测评、大模型X光测评等在内的测评和防御技术链条，并已运用于金融、政务、医疗等专业场景下的AI应用全流程。

常永波说，大模型落地门槛正在大幅降低，大量中小企业在模型安全治理方面的能力较弱，有些甚至不符合基本的合规要求。解决这些问题，需要监管的进一步引导和头部厂商的能力释放。

“我们现在已把‘蚁天鉴’的测评能力框架开源，将来也会把检测能力以及对风险的认知更多分享到平台上，它可以适配较多模型。希望我们提供的开放能力能帮助大模型行业持续健康发展。”王维强说，模型厂商离用户最近，可第一时间发现安全隐患，并通过和监管保持良性沟通互动，助力大模型安全落地。

清华大学长聘副教授李琦认为，大模型安全应用是一个新兴领域，研究和应用尚处于起步阶段。随着新的实践不断深入发展，相关技术也会持续升级，为建构大模型安全实践范式打造高价值参考体系。

人工智能治理是全球性问题。WAIC 2024开幕式上发布的《人工智能全球治理上海宣言》提出，高度重视人工智能的安全问题。宣言强调，以发展的眼光看问题，在人类决策与监管下，以人工智能技术防范人工智能风险，提高人工智能治理的技术能力。宣言呼吁，推动制定和采纳具有广泛国际共识的人工智能的伦理指南与规范，引导人工智能技术的健康发展，防止其被误用、滥用或恶用。

国内首个体育大模型发布

科技日报讯（记者**何亮**）记者7月12日获悉，国内首个体育大模型——上体体育大模型近日发布。大模型由上海体育大学与百度合作成立的上体—百度飞桨智慧体育技术创新中心研发，具有数据专、算法强、算力优、应用广等特点。

上体体育大模型包含体育文献、动作识别与技战术分析、多模态三个垂直大模型。其中，体育文献大模型通过学习国内外体育文献资料，能对体育问题进行专业系统解答；动作识别与技战术分析大模型能自动解析体育训练的视频与图像，有效输出人体姿态的解析结果及距离、速度、高度、角速度等量化指标，有助于深入分析生物力学；多模态大模型能有效支持学科交叉融合研究，其数据分析结果可用于比赛视频AI解说及个性化课程生成。

为方便使用，上体体育大模型在百度文心智能体平台支持下，推出上体体育大模型智能体。目前，大模型及其智能体已应用在体能训练、足球、羽毛球、网球等场景中。

据悉，大模型研发团队正在服务跳水、游泳、田径、体操、蹦床、攀岩等多支国家队的日常训练和巴黎奥运会备战工作，通过在训练场馆部署人工智能智算设备和应用系统，助力运动员提升训练质量，提高在国际赛场的竞争力。

南京：打造人工智能产业发展高地

◎本报记者 张 晔 实习生 普京文

每年打造30个标杆应用场景、统筹智能算力超6000P FLOPS（1P FLOPS等于1000万亿次浮点运算/秒）……这组规划数据绘出南京人工智能产业创新高地蓝图。

近日，南京市发布人工智能行动计划及政策措施“1+1”文件，即《南京市进一步促进人工智能创新发展行动计划（2024—2026年）》（以下简称《行动计划》）和《南京市促进人工智能创新发展若干政策措施》（以下简称《政策措施》），旨在通过政策引导、创新驱动、应用牵引，打造具有全国影响力的人工智能产业发展高地。

南京市是工业和信息化部批复建设的全国第9个、江苏唯一的国家人工智能创新应用先导区，具备雄厚的科技基础和人才优势，拥有发展人工智能产业的良好条件和广阔前景。

《行动计划》提出，到2026年，力争引进培育国内外先进水平的基础大模型1个，打造行业大模型20个以上，实现人工智能核心产业规模600亿元。为推动落实《行动计划》，《政策措施》提出支持算法创新突破、提升算力支撑能力、推动“人工智能+”应用创新示范和构建良好产业生态四方面12条措施。

在日前举行的2024南京人工智能产业发展论坛上，南京市“人工智能+”应用创新示范场景征集行动正式启动，深度挖掘“人工智能+”与各行各业融合创新应用场景。

打造南京人工智能产业创新高地，政府在行动，行业也在发力。在2024南京人工智能产业发展论坛上，南京经济技术开发区和中国信息通信研究院人工智能研究所签约共建大模型技术与应用服务平台。双方合作早已结出硕果，共建的江苏省唯一人工智能公共技术服务平台，近年来已对接服务百余家企业。此次双方在人工智能大模型领域进一步加强合作，将推动省级人工智能公共服务平台能力进一步拓展，为院地合作开启新篇章。

在此次论坛活动中，共有12个人工智能项目签约落地中国（南京）智谷，涵盖大模型、边缘计算等人工智能前沿领域。

全国电力领域AI大赛举行

科技日报讯（记者**罗云鹏** 通讯员**郑婕蓥 黄勇华 刘杰**）记者7月13日获悉，全国首个覆盖输电、变电、配电、安全管理等生产领域的人工智能（AI）大赛——南方电网2024年生产域AI算法应用竞赛决赛日前在广东深圳举行。大赛由国家发展改革委、国务院国资委指导，南方电网公司主办。

“电力生产领域的AI应用处于起步阶段。”南方电网公司输配电部副总经理章彬介绍，近年来，发电、输电、变电、配电、用电各环节数字化、智能化水平提升，海量应用场景产生。要更好识别分析这些场景，急需AI技术介入。

例如，利用人工手段进行导地线锈蚀、树障识别时，每张图片需花费两三分钟。而通过AI技术，只需80毫秒就能快速识别风险点。南方电网深圳供电局三级拔尖技能专家高德民介绍，对于导地线锈蚀、树障距离判别等难题，AI算法已经能完全满足生产需要。下一步，将进一步提升对电力复杂场景识别的准确率。

南方电网有关负责人说，未来将立足产业发展需要，举办电力生产AI行业级全链条、大规模、高质量赛事，促进电力AI产业发展，打造电力AI生态圈。

网络空间资产测绘：为安全防护“画出”实时“地图”

◎本报记者 张佳星

网络空间已成为继海、陆、空、天之后的“第五疆域”，对其空间布局进行摸排并形成“地图”，是维护网络空间安全的基础性工作。近日在2024全球数字经济大会上发布的DayDayMap全球网络空间资产测绘平台，能为用户提供全面、精准、实时的全球网络空间资产测绘服务。平台研发方远江盛邦（北京）网络安全科技股份有限公司（以下简称“盛邦安全”）董事长权小文告诉记者，随着数字经济发展，数据转化为无形资产，即便是拥有这些资产的机构也难以尽数掌握其分布状况，任何节点都有可能成为安全漏洞，因此进行网络空间资产测绘十分必要。

外围突破防不胜防

“互联网出现已经几十年，当前的网络空间和过去有天壤之别。”权小文解释，过去的网络安全策略是“防御”，只要守住自己的网站、数据库等即可。而现在，网络已融入现实中的各个领域，需要安全防护的范围不断延伸。比如，日益多样化且难以部署传统防御措施的物联网设备容易成为网络攻击焦点；随着数据中心、云服务等技术发展，网络资源分布呈现零散、随机趋势，外围突破防不胜防。

“数字经济的发展对网络安全提出更高要求。”权小文说，由于能源、化工等行业的设备、数据等资产实现了联网，若无有效保护，这些数据资产将面临安全风险。

我国近年来实施网络安全法、数据安全法、《关键信息基础设施安全保护条例》等法律法规，指导各单位进行实战化网络攻防演习，从各个层面加强网络安全建设。

“我们也会扮演‘蓝军’为相关单位提供攻防演练服务，即便对方防守严密，也会由于网络资产‘摸底’不完备而产生漏洞。”盛邦安全防护产品线高级总监聂晓磊说，这进一步表明，进行全要素网络资产测绘并形成“地图式”全局掌控，对安全防御、管理乃至整个建设运营工作非常必要。

合作打造灵敏“探针”

那么，怎样测绘虚拟且庞大的网络空间？如何全面掌握实时变动的网络空间状况？

权小文说，IPv4（互联网通信协议第四版）时代，通过“发信息包、接受反馈”等类似“地理信息实地勘测”的方式，可探测网络空间资产所在位置、匹配设备等信息，但这一方法在IPv6（互联网协议第六版）时代失效了。

在这样的背景下，盛邦安全和清华大学合作研制出DayDayMap全球网络空间资产测绘平台，聚焦全球IPv6网络空间测绘，攻克IPv6网络空间设备隐匿性强、探测效率低、探测成本高等难题，提出活跃IPv6地址探测方法体系，应用高效拓扑发现、网络安全策略遥测等先进技术。这相当于给平台装上了更灵敏的“探针”，使活跃地址发现速度提升30—911倍，端口探测效率提升136倍，显著降低探测成本。平台支持主被动资产测绘、资产类型识别等服务，当前发现110亿个活跃IPv6地址，覆盖超92%的全球路由前缀空间、28个行业、33大类、1100个子类，是目前识别IPv6资产最多的公开测绘平台。

“当前，网络空间‘地图’不断走向精准化、规范化。”权小文说，盛邦安全与相关机构联合发起的网络安全技术网络空间测绘数据交换格式编制工作正在进行。标准的形成以及合作单位的加入，将使网络空间“地图”在安全风险监管与预警、智慧城市数字资产感知与运营、IPv6资产发现与攻击面管理、网络空间挂图作战监测与处置等多个场景中得到进一步应用。

2024全球数字经济大会上，参观者在观看数字安全网络底座示意图。 陈晓根/视觉中国

星地异构移动性管理技术获新进展

科技日报讯（记者**崔爽**）记者7月13日获悉，中国电信研究院日前发布了基于标识的星地异构移动性管理技术试验成果。该技术旨在解决复杂星地异构网络的移动性管理难题，提升网络性能，优化用户体验。试验结果显示，相比传统异构网络间的切换，基于标识的星地异构移动性管理技术将网络切换完成时长从秒级降至百毫秒级，时延稳定性显著提升。

随着卫星网络和地面移动网络不断发展，天地一体、星地融合成为未来通信重要趋势。特别是在复杂的低轨巨型星座组网环境下，移动性管理面临网络拓扑动态时变和大尺度空间异构切换难题。试验验证了基于标识的星地异构移动性管理技术方案和原型系统的可行性和有效性。

据悉，试验由中国电信研究院联合北京交通大学，依托中国电信大科创装置开展。相关方案和原型系统有望应用于未来星地异构网络。