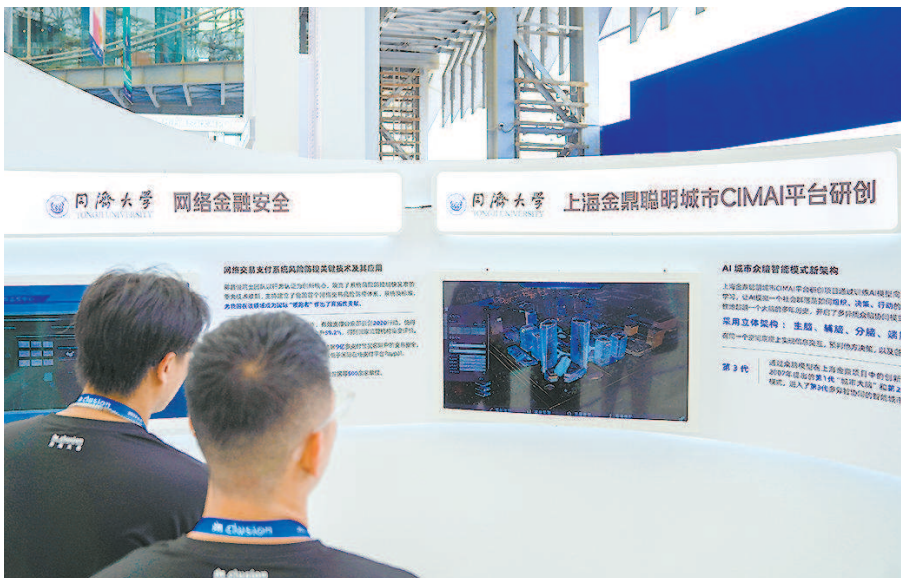
如何应对人工智能安全风险？

中国电信、蚂蚁集团、华为、百度等发起成立AI安全工作组

在昨天下午举行的“探索下一代安全智能”论坛上，世界权威国际产业组织“云安全联盟”(CSA)大中华区宣布成立“AI安全工作组”，中国电信、蚂蚁集团、华为、百度、火山引擎、西安电子科技大学、国家金融测评中心等30余家机构成为首批发起单位。该组织致力于共同解决AI技术快速发展所带来的安全难题。



■ 同济大学分享网络金融安全等方面的研究成果 本版摄影 本报记者 **陈梦泽**

云安全联盟大中华区AI安全工作组将由中国电信上海研究院、蚂蚁集团担任联席组长单位，召集联盟内人工智能相关的产线上下游企业、学校、研究机构、用户单位等共同研究人工智能安全前沿技术。

AI安全工作组将聚焦人工智能突出的安全风险问题，输出人工智能内生安全、衍生安全、助力安全等领域的白皮书、产业知识图谱、团体标准、系统框架、解决方案等，为整个行业提供一个清晰、系统的AI安全研究框架。根据计划，工作组先期将输出《AI安全白皮书》《大模型安全研究报告》《AI数据安全评估规范》等研究成果。

CSA(云安全联盟)创立于2009年，是世界领先的权威国际产业组织，共有成员单位1000余家。CSA致力于定义和提高业界对云计算和下一代数字技术安全最佳实践的认识和全面发展，凭借敏捷性、中立性、专业性被各界认可。CSA GCR(云安全联盟大中华区)是CSA全球四大区之一(其他三大区分别为美洲区、亚太区、欧非区)，立足于中国，作为国际桥梁连接世界，致力于构建国际数字安全的生态体系。此次AI安全工作组的成立，也将发挥国际桥梁作用，使我国AI安全研究成果在全球范围内产生影响。

从安全角度看，AI技术是把“双刃剑”。它既会对数据、业务和系统等层面构成威胁，也能为安全科技发展赋能。论坛与会嘉宾表达了一致的观点，并提出了建设性意见。

中国网络空间安全协会副秘书长徐倩华表示，这需要各行各业形成筑牢安全防护屏障的共识，注重培养顶尖安全技术人才，贯彻落实相关法律法规，共同提升数字安全防护能力，确保数字经济健康可持续发展。

云安全联盟大中华区主席李雨航院士也表示，人工智能技术的研发既不能因噎废食，也不能坐视安全不顾。优秀的中国AI方案应该成为全球的样板。

中国电信研究院副院长、云网基础设施安全国家工程研究中心常务副主任李安民认为，面对日益复杂的云网安全威胁，积极探索AI安全、Web3框架下的安全技术及其创新应用，才能帮助数据隐私、交易可信和网络稳定。

对于金融领域的科技创新与安全，国家金融与发展实验室副主任杨涛则表示，一方面要解决科技企业与金融机构建立“各司其职、风险自担”的合作边界问题，另一方面要完善金融科技生态与夯实基础要素。本报记者 **金志刚**

银发宣讲团妙语连珠讲述前沿技术

外滩大会探索助老新范式

当数字技术邂逅助老，会产生怎样的共鸣？这场数字峰会“人文温度”究竟有多少？从银发宣讲团现场讲述前沿技术，到hello老友亭2.0上新，蚂蚁集团蓝马甲行动和生态伙伴一起带来诸多助老创新，用温暖情怀融化“高冷”科技，帮助老年人跨越数字鸿沟。

老人给年轻人讲解前沿数字技术

“比如说你从外滩大会导航到杭州西湖，系统会帮你优化出来一条最便捷的跨城线路。”图计算展台上，蓝马甲银发宣讲团成员张益平正在给参会的年轻人介绍什么是“图计算”，说得通俗易懂点，图计算就是发现隐藏在事物背后的关联性，做出更聪明的选择，“我们老人自己先学会，再用过去的生活经验，把它们总结出来。我们老年人都能学会，可以给更多人信心。”

为助力银发族跨越数字鸿沟蚂蚁集团与上海老年大学联手，共同打造了一支蓝马甲银发宣讲团。团员从众多报名的学员志愿者中层层选拔，以蓝马甲志愿者的身份接受全方位专业培训，掌握了前沿科技的理论知识和实践操作能力，成为合格的银发宣讲员。他们在外滩大会上，用大家容易理解的方式，为往来观展的嘉宾讲解前沿数字技术。

“针对银发宣讲团，我们分别进行了隐私计算、图计算、可信AI等技术的培训，蚂蚁集团资深业务专家担任培训讲师，用通俗易懂的语言，告诉大家这些前沿科技是什么，用贴近生活的真实案例及场景展示这些技术对我们生活的帮助。”蚂蚁集团展会中心总经理、外滩大会负责人王文强向全上海的市民发出邀请，希望观众能够在黄浦江畔来一次全方位的前沿科技体验，“期待银发宣讲员们以专业讲解带领大家沉浸式感受科技力量”。

手绘助老说明书登上老友亭2.0

“摄像头在这里，如果要刷脸的话，看这里……”现场，一个红色的电话亭引起了记者的注意，和上海街头的电话亭不同的是，这个升级版的老友亭2.0多了一些新的功能，其中就包括了一份由蓝马甲志愿者、上海政法大学学生蹇丹手绘而成的使用指南。蹇丹告诉记者，在自己家门口发现老友亭，但是发现很多老年人其实不太能理解这些按钮如何使用，就设计了一份电话亭的“说明书”。

今年7月，蓝马甲联合上海电信发布第一期“Hello 老友亭N大功能介绍”，用简洁明了和口语化的图文表达，拆解老友亭中五项实用功能的操作步骤及注意事项，“如果变成电子版的，直接出现在电话亭的电子屏里，是不是能帮助到更多爷爷奶奶？”常年从事志愿工作，蹇丹说做就做，把设计图手绘出来以后，她联系了上海电信的工作人员，经过优化，即将陆续出现在街头的老友亭里。

蓝马甲“技术+服务”一直在身边

体验火眼金睛的VR游戏，和反诈盲盒一较高下，玩游戏可筛查是否有阿尔茨海默病风险……边走边看、边玩边探，外滩大会里助老反诈黑科技互动活动很多。

“这次外滩大会上，我们希望通过全方位的智能适老化设备、自助反诈游戏、AI大脑训练设备等，不仅能为老年人创造温馨的无障碍适老生活空间，更希望通过服务提升未来老年人的生活质量。”蓝马甲行动发起人王亦菲表示，科技并非无所不能，但是蓝马甲助老服务的初心一直不变，无论是银发宣讲团，还是青年志愿者，蓝马甲都会更加注重“科技+服务”模式的完善，注重志愿者线下服务的可触达性，“下半年，蓝马甲会继续通过进社区、大篷车下乡、助老公益集市等方式，联合更多的合作伙伴一起，通过‘面对面’‘手把手’的方式，走到老年人的身边，陪他们一起跨越‘数字鸿沟’。”

本报记者 **金志刚**

■ 蓝马甲“银发宣讲团”帮助参观者了解新科技

中科院院士何积丰：

为大模型安全设“紧箍咒”

模型的通用能力让其能够应用到人类生产生活的各个场景中，可谓“无孔不入”，也因此带来了新的安全隐私问题。如何解决这一隐患，中国科学院院士何积丰开出了他的“药方”：利用对齐技术为大模型设“紧箍咒”。

何积丰院士认为，大模型的安全问题主要是在未经同意的情况下，收集、使用和泄露个人信息。隐私问题既可能发生在训练过程中，也可能发生在使用过程中，而大模型的生成能力则让“隐私泄露”的方式变得多样化，造成隐私保护更加困难。

“为了应对这些问题，我们需要大模型对齐技术。”何积丰说，“对齐(alignment)”是指系统的目标和人类价值观一致，使其符合设计者的利益和预期，不会产生意外的有害后果。“如果把人工智能看作《西游记》里的孙悟空，那么‘对齐’就是唐僧的‘紧箍咒’。有了‘紧箍咒’，就可以保证技术不会胡作非为。”不过，对齐技术同样面临挑战。一方面，对齐的基础，人类的价值观是多元且动态变化的，需要保证大模型为人服务、与人为善；另一方面，大模型的有用性与无害性之间目标也不完全一致。如何对错误进行有效纠正，设好大模型的“紧箍咒”也是挑战。

本报记者 **杨硕**