开源活水 安全为堤

2026年世界互联网大会乌镇峰会观察

■ 记者 岳付玉

以后真能指望机器人养老吗？能把单位内部文件交给AI吗？AI骗你咋办？豆包、ChatGPT的闭源路线挺成功的，峰会主题为何聚焦“开源开放”？开源会不会带来更多风险？

带着一连串疑问，11月2日至5日，记者走进千年水乡，在2026年世界互联网大会乌镇峰会上寻找答案。

对这些话题感兴趣的，还有来自138个国家和地区的约1800位嘉宾，他们中很多人本身就站在AI大潮的潮头。他们的观点值得倾听。峰会的重要品牌活动——“互联网之光”博览会，则荟萃了54个国家和地区的千余项前沿科技成果，生动地“剧透”着“人机共生”的样子。

展台上，未来近得伸手可触；会场外，千年仍在小桥流水间。新与旧在水巷各安其位，又彼此成全。枕水乌镇，年年赴约。



蓝皮书里的两面

机器人逼近“ChatGPT时刻”，钓鱼网站制作只要5到10分钟

每年的乌镇峰会，都会发布两套蓝皮书：《中国互联网发展报告》和《世界互联网发展报告》。2026年的发布格外引人关注。全球人工智能正经历一场从量变到质变的跨越，数字经济进入智能化的全新发展周期，世界主要国家纷纷将人工智能纳入国家战略。

说到发展，蓝皮书给出的信号很积极。

蓝皮书介绍，2025年，我国数字经济核心产业增加值占GDP比重达10.5%以上，连续13年保持全球最大网络零售市场。数字基础设施同步下沉，全国95%的行政村已通5G，农村互联网普及率达69.5%。

具身智能产业快速兴起。2025年我国企业级智能体市场规模突破232亿元。具身智能市场规模约9150亿元，同比增长20.4%。宇树科技创始人王兴兴判断：2027年或成机器人行业“临界点”，具身智能已接近自己的“ChatGPT时刻”——在80%的陌生场景中，通过语音或文字指令，机器人能完成80%的任务。

人工智能的活跃度有一个指标很直观：我国日均词元调用量从2024年初的约1000亿，攀升至2026年3月的140万亿，两年多增长超过千倍。在蓝皮书统计期间，自2026年4月下旬以来，我国大模型周词元调用量已连续三个月居全球首位。

但蓝皮书并未只谈发展，它用相当大的篇幅指出“新型网络安全风险挑战加剧”。

攻击端的变化最直观。2025年人工智能赋能的攻击量同比增长89%，大多数威胁者已将人工智能整合到攻击行动中。生成式人工智能大幅拉低社会工程学攻击门槛、放大攻击效能。人工智能生成的钓鱼邮件在月度钓鱼尝试中的占比从不足5%激增至56%；制作一个钓鱼网站所需时间从原来的半小时到一小时，缩短至仅5到10分钟。

供应链是另一个突破口。2026年3月，人工智能模型代理框架LiteLLM（一种开源工具）在PyPI（Python 官方的软件仓库）遭遇供应链投毒，可自动窃取受害者机器中的密钥、云服务凭据等敏感信息。LiteLLM月下载量超过9500万次，波及大量下游人工智能项目。2026年4月，国家网络与信息安全信息通报中心发布预警指出，近期集中爆发多起供应链投毒攻击事件，攻击目标包括Apifox、LiteLLM以及JavaScript库Axios，涉及开源软件仓库和商用工具两大核心供应链场景。开发者终端、代码托管平台成为主要突破口，开发工具和开源组件一旦被攻陷，极易沿供应链链路大范围扩散——开源带来“活水”，也把“堤岸”问题摆在眼前。

蓝皮书还指出，APT攻击（即高级持续性威胁）向“精准制导”升级。2025年，APT-C-26（拉撒路）、APT-C-47（旺刺）等组织已广泛利用大模型进行深度伪造，生成虚假面试与会议邀请，推动钓鱼攻击向高仿真、交互式方向升级。

治理侧的回应也记录在蓝皮书里。我国将人工智能纳入国家网络安全法律体系，2026年4月，国家网信办部署开展“清朗·整治AI应用乱象”专项行动，重点整治未按规定履行大模型备案登记义务、人工智能数据投毒、生成合成内容标识落实不到位等问题。今年上半年全国各级网信举报工作部门、主要网站平台受理违法和不良信息举报1.09亿件。截至2026年8月底，累计有1112款生成式人工智能服务完成备案。

发展迅猛、乱象频发、治理提速，这是当下中国互联网生态的真实截面。把它们放在同一份报告里，本身就耐人寻味。

开源为何成主题

不争路线，争的是参与

开源，好比把菜谱全部公开，谁拿到都可以照着做、调整口味；闭源，就像饭店直接做好菜端给你吃，味道由店家把控，配方不对外公开。开源方便大家一起用一起改，闭源让使用者买成品省心但受限制。

——这是豆包和DeepSeek对开源和闭源给出的解释。前者是闭源，后者是开源。

在国内，DeepSeek、阿里千问、腾讯混元等都坚持走开源路径。像蚂蚁集团旗下具身智能公司蚂蚁灵波，其系列模型已走进斯坦福大学Open-WAM项目、韩国FastOPD、瑞士洛桑联邦理工学院等全球科研与工程体系。英伟达也将LingBot-World（灵波世界）集成至其开源世界模型推理库。

走闭源路线的豆包、ChatGPT做得也很好，说明闭源不是一条走不通的路。那么峰会为什么还要把“开源开放 共建共享——携手构建网络空间命运共同体”定为主题？

在主论坛上，中国工程院院士、之江实验室主任王坚问：“为什么要坚定选择开源？”

刘大一恒的回答没有停在“技术路线之争”上。这位阿里巴巴ATH事业群千问大语言模型项目负责人说：“开源让核心技术不只掌握在少数企业手中，还让全球开发者和生态共同参与创新；开源不仅是开放模型使用权，更是给予了全球每个人参与创造AI的权利。”他还讲了一个很具体的理由：世界上有很多问题可能头部公司并不关心，比如小语种、特殊垂类应用等，但有了开源，这些才有人去做。开源还可以赋能学术界，世界上有不少学者不缺乏想象力、创造力，缺的是算力，开源模式为他们提供了一个很好的平台。

三年多前，刘大一恒亲手完成了千问第一代7B模型的训练并开源，此后，该系列陆续开源了460多款模型，全球下载量超过40亿次。

北京稀宇极智科技有限公司首席架构师冯雯以亲身体验为例，她说：“我们把模型开源之后，经常会出现意想不到的好事。芯片厂商、开发者、内容创作者都来做适配，还把自己做模型训练的工具发到开源社区中。开源让我们看到，一家公司能做的其实有限，但开源可以把边界推得更远。”

阿里巴巴集团董事会主席蔡崇信认为，开源是AI走向普惠的重要路径。只有开源才能让全球开发者生态真正繁荣，新兴经济体开发者不必从零起步，可直接基于开源模型打造本土AI应用和生态。

王坚紧接着抛出了第二个问题——安全。“把模型交出去，换回来的是更快的迭代和更大的生态。但技术交到更多人手里，隐私与安全问题也无法回避。”他说。

刘大一恒没有回避该问题。他承认：“相对而言，开源模型确实面临一些额外的安全挑战。”第一，开源模型具有不可撤回性，模型一旦开源就会被大量传播，如果它有问题，确实无法像闭源模型一样实时在线修复。第二，开源模型缺少外部防护屏障，它的安全性很大程度上依赖后训练内生安全能力，第三方拿到模型权重后，可以想办法绕过内置安全护栏。不过，在他看来，安全单靠限制模型开放难以从根本上解决，还需从模型之外通过法律法规、技术规范协同配套。

DeepSeek研究工程师高熙说得更实在：“不开源同样存在安全风险，恶意可能来自人，也可能来自AI。”国际人工智能治理协会主席约翰·希金斯则把责任主体点明：“保障开源安全，已成为全社会需要共同承担的责任。”

开源与闭源在安全维度上各有短长：开源的风险突出体现为风险易扩散、模型不可撤回；闭源的隐患集中在黑箱不透明、权责高度集中。治理要做的不是选边站，而是让两条路线都在可审计、可追责的框架里运行。

记者在采访中感受到，峰会聚焦“开源开放”，不是断言“开源优于闭源”，而是在表达公共立场：AI不能只作为少数公司的成品交付给世界，它还需要一条让更多国家、更多开发者、更多研究者参与创造的通道。闭源的成功更多体现为企业的成功；开源开放要回答的是“全球大多数人能不能参与、能不能不被甩下”。峰会表达的是发展与治理姿态：在“构建网络空间命运共同体”下，技术扩散的普惠性、参与性、公共品属性，需要制度性出口。开源是其中之一。

智能体治理之难

不仅在于“能不能干”，更在于“能不能被托付”

在乌镇峰会的首场分论坛上，智能体注册解析平台正式发布。国家层面推动共建这个平台，重点解决身份编码、跨域解析和可信互通。作为先导验证，“智联浦江”行动计划稳步推进，支付宝晓政、芝麻信用等首批服务已接入运行平台，并完成与其他平台智能体的可信连接。

眼下，AI应用正从“会回答”走向“能办事”。但这个转变把治理问题从“模型说了什么”推到了“智能体做了什么”。开源开放又把模型、工具链、接口和编排框架摊到了更多主体手里：一个智能体可能调用开源模型、第三方工具和跨平台服务，身份是谁、权限多大、行为如何审计、责任由谁承担，很难再由单一厂商闭环。开源开放走到“行动”层面后，治理规则也必须同步走向开放、可审计、可追责。

智能体虽然开始感知、决策并代表用户行动，但与会者认为，其规模化落地仍面临三道坎：用户能否知晓并及时撤回智能体权限；服务方能否确认调用的智能体是谁、代表谁行动；不同平台的智能体能否彼此找到并调用服务。这三道坎，在开源开放的生态里会被进一步放大：调用链更长、参与主体更多、责任边界更分散。

从“能说”到“会做”，要求更高。曙光云副总裁徐良指出，用户顾虑的不只是AI够不够先进，更在意安全是否可控、数据是否安全、成本是否可测、管控是否落地。AI要进入生产系统，安全不是事后选项，而是入场门票。

说到“入场门票”，博览会上也有生动注脚：曙光云新推的OneBox智能体协作平台（一种政企智能体协作平台），把模型、知识库和业务应用私有化部署在单位自有环境中，敏感数据不出域；腾讯WorkBuddy（一款 AI 办公助手产品）以沙箱限定Agent行为，高危操作拦截、越界请求确认；蚂蚁APASS（智能体可信身份基础设施）从身份和授权切入，让授权“看得见、收得回”……各家形态不同，逻辑却一致：可信AI不是靠用户自觉，而是把权限、数据、审计和防御做进系统。

从管“语言”到管“行动”，治理挑战更进一步。中国工程院院士云晓春在论坛上说得很犀利：具身智能把AI从虚拟世界延伸到物理世界，而物理世界造成的伤害往往“不可接受、不可逆”。能力越强，安全边界越不能模糊；治理关口必须前移，不能等模型上线后再打补丁。中兴通讯董事长方榕直接点出：“AI跑得越快，越要系好‘安全带’。”

“智能体规模化的上限，不仅在于‘能不能干完一件事’，更在于‘能不能被托付一件事’。”蚂蚁集团副总裁陈亮说，“智能体时代需要构建信任基础设施，让身份、授权与责任进入每一次连接。”这不仅是企业产品能力的竞争，更是智能互联网发展的公共基础设施问题。

智能体注册解析平台解决的是“连得上”。但连上之后，边界谁定、责任谁担，仍是难题。开源开放提供的不是现成答案，而是一种治理要求：规则要公开，才能被检验；责任要可追溯，才能被托付。

乌镇峰会主论坛发布的《前沿人工智能安全与治理2026：风险认知与协同治理》报告引入“治理就绪度”评估，覆盖24个国家，结果显示全球AI安全与治理的制度建设仍存在显著缺口。与会者在此基础上提出推动模型安全评测协同、加强风险事件通报与联合处置等五大建议。报告发布人、世界互联网大会人工智能专委会主任委员曾毅的一句话让人印象深刻：“每一个国家在人工智能发展面前和安全挑战面前都是小学生。”

人的尺度

技术再强，也要接住具体的人

技术治理越往深走，越绕不开一个问题：这些能力最终为谁所用？

走进“互联网之光”博览会，答案先以场景铺开：24小时药房里机器人在值守，健康管理有AI私教和营养师，辅导作业有能把立体几何“化抽象为具象”的AI黑板，出行有动动口就能规划订票的智能体手机，腿脚不便的人戴上租来的外骨骼走台阶如履平地，居家养老可以让机器人按摩、做饭……这些场景共同指向一件事：“人机共生”不是概念，是具体的人被技术接住的瞬间。

腾讯集团高级副总裁奚丹把这条逻辑说得很清楚：“人是AI发展的尺度。发展AI，核心是放大人的价值和作用，让AI成就人，把技术的温度送到每一处边缘角落。”这话由一家把AI注入社交、支付、就业、公共服务的公司来说，不是抒情，而是产品逻辑——AI的价值不只在峰值能力，也在边缘覆盖。

丁香园创始人李天天给医疗AI划的底线是：医疗行业低容错率决定了创新必须稳健迭代，AI仅提供循证依据，不替代医生诊断——不是AI能不能做，而是AI做了之后谁担责。其实何止医疗，自动驾驶、金融，莫不如此。

蔚来创始人李斌用数据说话：蔚来智能驾驶累计安全行驶超91.4亿公里，人车共驾平均安全里程达人类驾驶的8倍，AI为用户规避或减少超907万次潜在事故。技术进化的归宿应当是解放人，而不是淘汰人。

开源开放把参与创造的门槛降下来，但技术往哪里走、为谁所用、由谁担责，最终仍要回到人的尺度上。

