畅谈

让人工智能既“聪明”又可信

■ 大会开幕式现场 本报记者 陈梦泽 摄

本报记者
郜阳 叶薇 马亚宁 易蓉

上午10时，2024世界人工智能大会暨人工智能全球治理高级别会议在上海拉开大幕。在全体会议上，清华大学苏世民书院院长、清华大学人工智能国际治理研究院院长薛澜，上海人工智能实验室主任、首席科学家周伯文，黑石集团董事长、首席执行官兼联合创始人苏世民，微软原执行副总裁、美国国家工程院外籍院士沈向洋，图灵奖得主姚期智等专家学者带来精彩见解。

上海人工智能实验室主任周伯文：探索人工智能45°平衡律

在2024世界人工智能大会全体会议上，上海人工智能实验室主任、首席科学家，清华大学惠妍讲席教授周伯文发表主旨演讲，生动诠释“人工智能45°平衡律”。“当前，以大模型为代表的生成式人工智能快速发展，但随着能力不断提升，模型自身及其应用也带来了一系列潜在风险的顾虑。”他说。

从公众对AI风险的关注程度来看，首先是数据泄露、滥用、隐私及版权相关的内容风险；其次是恶意使用带来伪造、虚假信息等相关的使用风险；当然也诱发了偏见歧视等伦理相关问题；此外还有人担心：人工智能是否会对就业结构等社会系统性问题带来挑战。在一系列关于人工智能的科幻电影中，甚至出现AI失控、人类丧失自主权等设定。这些风险已初露端倪，但更多是潜在风险，防范它们需各界共同努力，需要科学社区作出更多贡献。去年5月，国际数百名AI科学家和公众人物签署公开信，表达对AI风险的担忧，并呼吁应该像对待流行病和核战争等其他大规模的风险一样，把防范人工智能带来的风险作为全球优先事项。

周伯文认为，有这些担忧的根本原因是目前的AI发展是失衡的。他向大家展示了一张坐标图——横轴是AI技术能力的提升，呈现指数级增长；纵轴是AI的安全维度，典型的技术如红队测试、安全标识、安全护栏与评估测量等，呈现零散化、碎片化，且后置性的特性。“总体上，我们在AI模型安全能力方面的提升，还远远落后于性能的提升，这种失衡导致AI的发展是跛脚的，不均衡的背后是二者投入上的巨大差异。”周伯文指出，“对比一下，从研究是否体系化，以及人才密集度、商业驱动力、算力的投入度等方面来看，安全方面的投入是远远落后于AI能力的。”周伯文说，可信AGI需要能够兼顾安全与性能，人们需要找到AI安全优先，但又能保证AI性能长期发展的技术体系，“我们把这样一种技术思想体系叫作‘AI-45°平衡律’”。

AI-45°平衡律是指从长期的角度来看，要大体上沿着45°安全与性能平衡发展，平衡是指短期可以有波动，但不能长期低于45°，也不能长期高于45°——这将阻碍发展与产业应用。这个技术思想体系要求强技术驱动、全流程优化、多主体参与以及敏捷治理。

周伯文介绍，实现AI-45°平衡律也许有多种技术路径，上海人工智能实验室近期在探索一条以因果为核心的路径，并将其取名为：可信AGI的“因果之梯”，致敬因果推理领域的先驱——图灵奖得主朱迪亚·珀尔。可信AGI的“因果之梯”将可信AGI的发展分为三个递进阶段：泛对齐、可干预、能反思——“泛对齐”主要包含当前最前沿的人类偏好对齐技术；“可干预”主要包含通过对AI系统进行干预，探究其因果机制的安全技术；“能反思”则要求AI系统不仅追求高效执行任务，还能审视自身行为的影响和潜在风险，从而在追求性能的同时，确保安全和道德边界不被突破。这个阶段的技术，包括基于价值的训练、因果可解释性、反事实推理等。

周伯文表示，目前，AI安全和性能技术发展主要停留在第一阶段，部分在尝试第二阶段，但要真正实现AI的安全与性能平衡，我们必须完善第二阶段并勇于攀登第三阶段。沿着可信AGI的“因果之梯”拾级而上，“我们相信可以构建真正可信AGI，实现人工智能的安全与卓越性能的完美平衡”。

清华大学人工智能国际治理研究院院长薛澜：携手合作弥合数字鸿沟

2015年，《联合国2030年可持续发展议程》提出全球要历时15年让全球的发展进入到可持续发展轨道。去年联合国有关会议透露，完成目标的情况不容乐观。“最近有分析表明，人工智能对可持续发展的169个子目标中的134个有积极促进作用，也有少部分会产生不利影响。要从宏观角度看待人工智能潜在的积极和不利影响。”清华大学人工智能国际治理研究院院长薛澜指出，如何推动人工智能的健康发展，尽可能让它收益最大化，把风险降到最低，中国这些年已建立相对完整的体系。

产业应用方面有一系列法律法规去推动人工智能合理发展；针对算法算力以及数据治理也有规则框架；在具体的场景应用上，也出台了具体规则。“我们构建了一个多维度、多层次、多领域、多举措的体系。”薛澜说，我们要非常重视国际社会的人工智能安全问题，通过多种途径建立国际交流，加强政府间的多边对话机制，用科学的力量助力国际机制的完善，促进政府、企业推动人工智能健康发展。

第78届联合国大会7月1日协商一致通过中国主提的加强人工智能能力建设国际合作决议。薛澜认为，国际社会亟需携手合作打破各种壁垒，共同推进全球人工智能治理。

黑石集团董事长苏世民、索奈顾问及投资公司董事长乔舒亚·雷默：AI+投资带来新思考

上午，黑石集团董事长、首席执行官兼联合创始人苏世民先生和索奈顾问及投资公司董事长、首席执行官乔舒亚·雷默先生的一场高端对话，给AI+投资带来新思考。

在雷默看来，人工智能颠覆着传统公司的价值，有的公司因AI一夜暴富至几十亿的资产。作为商业投资人，苏世民从去年开始就观察到，有一些企业在AI融合方面做得很棒，但也有一些企业好像不太适应。在寻找潜在投资机会时，AI已经成为很好的分析帮手。“投资界里有一句谚语，没有勇敢的老年人”，因为成功的投资人总是对可能出错的事情保持警惕。人工智能正在为这种投资的警惕性，施展很好的理解力和判断力。“AI已经成为我们看待事物的一个重要的一种方式。”苏世民说。

令雷默十分感兴趣的是，苏世民作为AI学术研究的重要捐资者，为何突然选择了AI领域？“我没有选AI，是AI选了我。”苏世民表示，人工智能在深入理科研究的同时，也需要与人类的情感相结合，并找寻到“理智与情感”的结合方向。未来10年至20年后，AI必将深刻影响人类的发展和未来，前沿科技与人文社会的有机协同，从现在开始就值得思索和探究。“未来，比仅仅资助一位教授研究项目更有意义的是，AI的产业发展和大学基础研究的有机结合。”苏世民说。

图灵奖得主罗杰·瑞迪、曼纽尔·布鲁姆、姚期智等：让聪明人打磨“AI的双刃”

“我们看到了人工智能带来的巨大能量，也为未来感到隐忧，如何能够实现平衡，是一个深刻的问题。”全体会议上，3位图灵奖得主罗杰·瑞迪、曼纽尔·布鲁姆、姚期智和原微软执行副总裁、美国国家工程院外籍院士沈向洋开展对话，这场“巅峰师徒”圆桌聚焦“AI的双刃”。

姚期智认为当前人工智能面临三种风险，既有原来互联网技术风险的扩大，也有失业等颠覆社会结构带来的社会风险，甚至还面临颠覆性力量过大带来的生存风险。“作为科学家，我们更关心第一种和第三种风险。”他解释，互联网信息技术已经带来数据管理的现实困难，AI会将这种困难扩大百倍。他认为，所有风险需要社会各界共同参与着手应对。

瑞迪则认为更应该聚焦人工智能对效率提升的研究，“试想，未来每个人的生产能力至少翻10倍，必然带来更大的社会经济价值”。而如何令每个人的工作效率都能得以提升，是研究者需要关心和解决的问题。

布鲁姆近几十年的研究都聚焦“意识”，他与合作者研发“有意识的图灵计算机CTM”即“有意识的人工智能技术”。这样的AI是否更聪明？他举例来让大家品味答案：“CTM没有中央决策者，让每个人把重要信息放到前台来。一家企业的首席执行官实际上并不可能了解所有员工的能力，但是他一个人将替企业做出决策。”

智能会展『小秘书』升级

“嘿，如果我从新闻中心出发，想去和人工智能大会上的人形机器人碰个头。你能帮我规划一个参观路线吗？”今天，上海市人工智能行业协会携手技术合作伙伴，推出超级会展智能体——WAIC AI Agent。这个又懂AI又“懂你”的智能会展助手是个全能的“小秘书”，正让2024年世界人工智能大会（WAIC）参会体验变得更轻松有趣了。

打开WAIC 2024智能会展的在线小程序，你就能看到WAIC AI Agent的卡通形象“威客兄弟”在云端朝你微笑招手。“嘿，想参加大模型的论坛，能给我安排一下？”热情的威客兄弟会迅速给出回应：“当然可以！您对大模型感兴趣，我为您推荐7月5日9时开始的‘可信大模型助力产业创新发展’论坛……”记者体验发现，WAIC AI Agent基于前沿的大模型技术，正为嘉宾和观众带来全新的便捷服务。它是个可移动的“百科全书”，还能成为你的“参会行程定制师”。只需简单一问，它就能告诉你论坛的名称、时间、地点，甚至还能帮你找到最佳的路线。

WAIC AI Agent也是逛展小能手。只要输入展位名称或感兴趣的类别，它就能实时提供最佳逛展路线规划，为你轻松指路，甚至它可以通过拍照上传的展台标识，识别并提供展商信息，简直是路痴的福音。记者向它提出：想去看一看展会上的人形机器人。它就为记者设计了一条不走回头路的路线，并详细罗列了各个人形机器人所在的展位编号。“为了帮助您从新闻中心出发找到有人形机器人的展台，以下是您的展位规划描述内容及展位顺序：首先，您可以从新闻中心出发，前往H1展区的a区，那里有多个展位展示各种科技产品。接着，您可以依次参观H1展区的b区、c区和d区，最后前往H2展区，那里也有许多与人形机器人相关的展台……”

本报记者 马丹