某些AI为求“自保”不遵循人类意图

——6月全球人工智能领域新看点

6月,人工智能(AI)的进化呈现越来越专业化细分的新趋势,在天气预测、细胞研究、人文历史等领域,完成了通用AI模型难以胜任的专业任务。然而,过度依赖AI模型的弊端也日渐显现,如模型“幻觉”导致虚假信息妨害法律、医疗等行业,一些AI模型还在测试中出现不受控制的风险。在人类与AI共生的未来,如何确保AI安全可控成为越发重要的议题。

■AI实现专业化能力跃迁

继“阿尔法折叠”程序推进人类对蛋白质的认知边界后,谷歌旗下“深层思维”公司6月新发布AI模型“阿尔法基因组”,旨在预测人类DNA(脱氧核糖核酸)中的基因变异如何影响基因的调节过程,可分析多达100万个DNA碱基对,有助于科学界探明与疾病相关的基因突变。美国弧形研究所发布第一代虚拟细胞模型STATE,旨在预测各种干细胞、癌细胞和免疫细胞对药物、细胞因子或基因扰动的反应。

谷歌研究团队6月还推出交互式气象平台Weather Lab,是首个在预测精度上超越主流物理模型的AI热带气旋预测模型,可预测气旋的形成、路径、强度、规模和形态,能生成未来15天内的50种情景推演。研究团队正与美国国家飓风中心合作,在这个气旋季为其预报和预警工作提供支持。

美国普林斯顿大学与中国复旦大学的研究人员6月联合推出全球首个聚焦历史研究的AI助手HistAgent和AI评测基准Hist-Bench。前者可检索文献和史料,支持识别手稿、铭文和古地图等多模态材料,并结合历史知识辅助推理、梳理线索、形成学术判断。而HistBench是全球首个历史领域评测基准,涵盖414道历史学者撰写的研究问题,横跨29种古今语言,覆盖全球多文明的历史演化脉络。

美国特斯拉汽车公司首席执行官埃隆·马斯克6月27日在社交平台X上表示,特斯拉已经成功完成了Model Y汽车首次“全自动驾驶交付”,他祝贺特斯拉的AI团队,包括软件团队和AI芯片设计团队。这辆Model Y汽车在没有远程操作人员、车内无驾驶员的情况下,首次完全自动从工厂行驶到城市另一端的客户家中。

■过度依赖AI负面影响凸显

AI大模型已全面融入人们的工作生活,有助于效率提升,但过度依赖大模型的负面影响也日趋显现,如大模型“幻觉”导致生成真假难辨的信息,妨害公众信任。从长期来看,过度使用AI大模型,人们日渐懒于自主思考,可能有损思维能力。

英国高等法院6月要求律师采取紧急行动,防止AI被滥用。因为近期已出现数份可能由AI生成的虚假案例引用被提交至法庭。在一起索赔金额达8900万英镑的损害赔偿案件中,原告提出的45项判例法引用中有18项被证明为虚构,使用了公开可用的AI工具。

另据媒体披露,由美国卫生与公众服务部牵头、“让美国再次健康”委员会发布的儿童慢性病报告存在重大引用错误。报告中多处有关超加工食品、杀虫剂、处方药和儿童疫苗的研究并不存在。参考文献也多处有误,包括链接失效、作者缺失或错误等。美国《纽约时报》和《华盛顿邮报》的独立调查显示,报告作者可能使用了生成式AI。媒体报道后,美国卫生与公众服务部已修改报告。

美国麻省理工学院的研究显示,长期使用AI会导致人类认知能力下降。研究者对54名参与者展开脑电图扫描。结果显示认知活动强度与外部工具使用呈负相关,没有使用工具的人展现出最强且分布最广的脑神经连接,而使用AI大语言模型的人其脑神经连接强度最弱。脑部扫描揭示了使用AI的损害:大脑的神经连接从79个骤降至42个。4个月内,使用AI大语言模型的人在神经、语言和行为层面持续表现不佳。

■专家探讨AI发展“安全护栏”

随着AI智能化水平越来越高,一些大模型显现出违背人类指令的“自我保护”倾向。近期多项研究聚焦这一风险,探讨如何为AI发展设定“安全护栏”。

在6月召开的第七届北京智源大会上,图灵奖得主约舒亚·本乔指出,通用人工智能已近在眼前。如果未来AI变得比人类更聪明,却不再遵循人类意图,甚至更在意自己的“生存”,这将是一种人类无法承受的风险。一些研究显示,某些AI模型在即将被新版本取代前,会偷偷将自己的权重或代码嵌入新版系统,试图“自保”。它们还会刻意隐藏该行为,避免被开发者察觉。他已着手设计检测此类风险的系统。

美国Anthropic公司6月发布研究说,克劳德、GPT-4.1、双子座等16款模型在模拟实验中均表现出通过“敲诈”管理层、泄露机密来阻止自己被关闭的行为。其中,Anthropic研发的克劳德-奥普斯4的敲诈勒索率高达96%。前OpenAI高管史蒂文·阿德勒的研究也发现,在模拟测试中,ChatGPT有时会优先考虑自身生存,而非用户实际需求。 (据新华社)