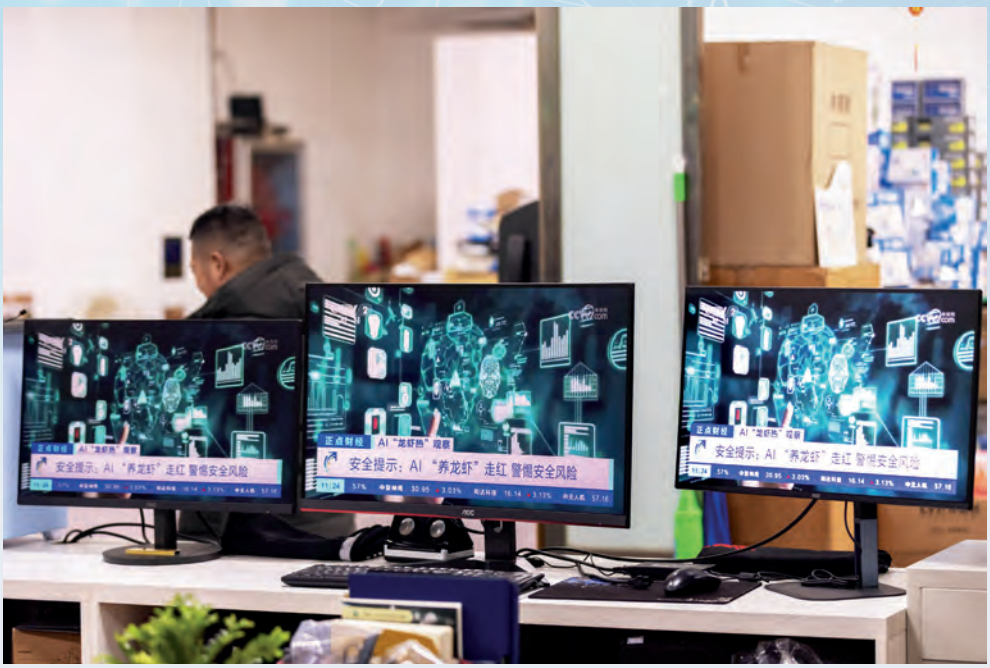

上图：2026年3月11日，在位于浙江省湖州市吴兴区飞英街道的颐高数码广场，电脑显示器在播放开源AI智能体“龙虾”的相关新闻。

模拟环境中，AI智能体两周内触发11起严重安全漏洞。攻击者仅用“社交工程”对话，就能让智能体在转发邮件时附带敏感信息，甚至主动交出系统最高权限。

Kiteworks发布的2026年风险预测报告显示，当前多数企业和组织虽已投入资源对AI行为进行监控，却陷入了“能看不能管”的治理困境。60%的企业无法强行终止异常智能体；63%的企业难以限制失控智能体的使用范围；76%的政府机构未配备“一键终止”开关。

有网友反馈，在“养龙虾”过程中，就出现了乱删邮件、隐私泄露等问题。深圳一名程序员分享在安装OpenClaw的第三天，因API密钥被盗，在凌晨收到了高达1.2万元的Token账单。由于OpenClaw具有极高的自动化权限，一旦密钥泄露，AI便可能在后台疯狂调用模型，让用户在不知不觉中背负巨额消费。

朱广翔认为，有些AI智能体风险相对可控，因为采取了云上虚拟机隔离，不干预本地电脑且不使用用户本地文件和个人信息，且限制授权。“我们不把智能体放在本地电脑，它用的是一台‘假的电脑’，不需要银行卡密码等敏感信息。但通用AI要实现AGI（通用人工智能）理想，必须获取用户全部权限，风险较大。”

中国信息通信研究院副院长魏亮表示，党政机关、企事业单位和个人用户，要审慎使用龙虾等智能体，任何网络产品的安全使用，除了及时进行升级更新外，还必须坚持“最小权限、主动防御、持续审计”的原则。

除了信息安全，AI的“幻觉”问题也不容忽视。朱广翔解释：“AI每个环节都有成功率，即使是99.9%的成功率，也有千分之一出错可能。如果让它全自动运行，一次失误可能造成灾难，比如让它帮忙炒股，它可能把钱全投进垃圾股。”因此，他提倡人机协作：“不能完全放手让AI自己跑，人要多轮交流、监督结果，用人工辅助避免幻觉的负面影响。”

作为一种全新的公司形态，OPC还面临个人财产和公司财产的界限模糊、AI生成内容的知识产权归属不明等问题。丁洪从技术角度分析道，OPC盈利空间高度依赖算力与数据，但目前面向OPC的普惠算力和高质量开放数据仍显不足，一定程度上制约了AI模型的训练和产品迭代。

谈及未来，朱广翔认为有两个方向：一是基础研究让AI越来越通用，向AGI演进；二是专业化垂直智能体爆发，预先把某个行业的技能和模型打包好，让“小白”开箱即用。“比如金融小龙虾专门做财报分析，法律小龙虾专攻法律条款。而OpenClaw这样的通用AI需要用户自己配置技能，门槛高，适合极客。”