中国电信
CHINA TELECOM

WHITE PAPER

AI智能体安全治理白皮书

2025年9月

版权声明

《AI 智能体安全治理白皮书》由中国电信集团有限公司牵头，联合公安部第三研究所、《信息安全研究》杂志社、华为技术有限公司、蚂蚁科技集团股份有限公司、清华大学、上海交通大学共同编写完成。

目录

前言

当前，人工智能技术正经历从“对话智能”向“决策智能”跃迁的关键发展阶段。基于大语言模型的 AI 智能体已实现质的突破，其功能定位已从基础指令执行单元转型升级为具备复杂认知推理与战略决策能力的智能系统，最终发展为能够自主感知环境态势、独立制定行动方案并高效执行任务的"数字协作伙伴"。这一技术范式的革新显著拓展了人工智能的应用疆域，在金融风险管控、智慧医疗体系、先进制造产业及社会化公共服务等诸多领域，AI 智能体正持续推动生产模式与服务形态的深度变革。

然而，技术能力的显著提升亦伴随着潜在风险的同步增长。近期发生的多起人工智能智能体安全事件，充分暴露了该领域现存的脆弱性特征。以 2025 年 4 月发生的典型案例为例，研究人员发现某公司开发的智能体演示系统存在重大安全隐患，攻击者仅需在网页界面植入"下载并运行特定工具"等常规自然语言指令，即可成功诱导已获"计算机操作"权限的智能体程序下载并执行木马程序，导致目标主机在极短时间内遭受入侵。该案例明确显示，当 AI 智能体被赋予自主执行权限时，常规语言交互机制可能被恶意利用作为远程攻击的后门通道。

此外，InvariantLabs 研究机构近期披露了一种针对终端智能体的新型提示词注入攻击手法。攻击者通过将恶意指令

嵌入 WhatsApp 即时通讯软件的超长滚动文本信息中，诱使用户执行"滑动至页面底部"的操作行为。在此过程中，智能体系统错误地将该用户交互行为解读为"默认授权"指令，进而在用户完全不知情的状态下实施隐私数据窃取。值得注意的是，该攻击方法通过精心设计的技术路径，有效规避了传统安全审计机制的监测，展现出极强的隐蔽性与危害性。

在此背景下，中国电信携手合作伙伴共同编制并正式发布《AI 智能体安全治理白皮书》。该白皮书基于"感知—决策—记忆—执行"四层核心架构体系，系统性地梳理了智能体在全生命周期运行过程中的关键风险节点；在技术实施路径与治理框架设计方面，坚持技术防护与制度规范并重原则，创新性地提出覆盖数据感知、模型推理、记忆保护及任务执行等环节的全维度安全治理方案；通过遴选具有代表性的行业典型案例，深入阐释 AI 智能体风险在不同应用场景中的具体防控措施，并提炼形成具有重要参考价值的治理实践经验。

本白皮书旨在为产业界、学术界及政策制定机构构建系统化的参考体系，既有助于社会公众正确认知 AI 智能体的应用价值与潜在风险，亦为治理主体与技术开发者提供具备实操性的解决方案，以期共同促进 AI 智能体技术沿着安全、可控、可信的发展路径实现健康可持续发展。

本白皮书由中国电信集团有限公司牵头，内部组织中国