

人工智能安全治理 白皮书（2026）

中国联合网络通信有限公司

华为技术有限公司

中兴通讯股份有限公司

北京百度网讯科技有限公司

奇安信科技集团股份有限公司

三六零数字安全科技集团有限公司

清华大学

北京邮电大学

中国科学院大学

浙江大学

南开大学

2026 年 9 月

版权声明

本报告版权属于中国联合网络通信有限公司研究院，并受法律保护。转载、摘编或利用其他方式使用本报告文字或者观点的，应注明“来源：中国联通研究院”。违反上述声明者，本院将追究其相关法律责任。



目录

一、AI 前沿技术发展	3
1.1 多模态大模型技术	3
1.2 智能体技术	5
1.3 具身智能技术	8
二、AI 安全治理体系	10
2.1 AI 模型与算法安全	11
2.2 AI 智能体安全	12
2.3 AI 数据基础设施安全	13
2.4 AI 算力基础设施安全	14
2.5 AI 安全运营	15
2.6 AI 监督与管理	16
三、AI 安全治理风险分析与挑战	17
3.1 AI 模型与算法安全风险	17
3.2 AI 数据安全风险	20
3.3 AI 应用安全风险	22
3.4 AI 应用场景安全风险	33
四、AI 安全治理政策与标准现状	39
4.1 AI 安全治理法律法规与政策现状	39
4.2 AI 安全治理标准规范现状	43
五、AI 安全治理技术解决方案与案例	52
5.1 中国联通人工智能安全治理一体化解决方案	52
5.2 百度智能体安全解决方案	55
5.3 基于中兴 AIOS 的智能体安全治理方案	57
5.4 华为智能体安全治理方案	59
5.5 360 全模态大模型安全卫士 R3.0 解决方案	61
5.6 奇安信智能体安全解决方案	63
六、AI 安全治理发展建议	65
缩略语	68
参考文献	69

前 言

过去一年，多模态大模型和智能体技术快速发展，AI 应用规模持续扩大、应用场景不断拓展。多模态大模型能够综合处理文本、图像、音频、视频等异构信息，具备更强的多模态感知、理解与内容生成能力，同时在任务规划、工具调用编排等方面的能力持续增强。以大模型为驱动的智能体能够围绕目标分解任务、组合工具，并通过与环境交互调整任务执行，从而完成复杂任务。智能体技术推动了人工智能技术能力由内容生成向复杂任务执行延伸，扩展了模型能力的边界。

基础模型能力的提升和智能体应用技术的快速发展，在提高任务执行效率的同时，也使人工智能风险由内容层面延伸至现实行为。多模态大模型可能因模态语义冲突、理解偏差产生错误或者幻觉推断；其内容生成能力还可能被用于生成不安全的多模态内容，增加了内容溯源和风险处置的难度。当多模态大模型作为核心基座赋能自主智能体后，这些复杂的不安全内容会被进一步转化并放大为具有实质性危害的不安全操作。这种从“内生缺陷”到“复杂内容风险”再到“实质性操作风险”的传导链条，使得传统以单一模态输出审核为核心的安全治理框架失效。因此，既要防范由基础模型架构、训练数据、安全对齐算法等技术要素的缺陷、脆弱性所引发的 AI 内生安全风险，也要防范人工智能在部署、运行和应用过程中引发、放大或外溢的 AI 衍生安全风险。

本白皮书聚焦多模态大模型、智能体、具身智能等前沿技术发展，深入分析 AI 技术发展面临的新形势、新风险与新挑战，并对 AI 安全治理体系进行全方位升级，推动我国人工智能朝着有益、安全、公平方向健康有序发展。

本白皮书由中国联通研究院主笔，中国联合网络通信有限公司、华为技术有限公司、中兴通讯股份有限公司、北京百度网讯科技有限公司、奇安信科技集团股份有限公司、三六零数字安全科技集团有限公司、清华大学、北京邮电大学、中国科学院大学、浙江大学、南开大学联合编写。

编写组成员（按姓氏汉语拼音字母顺序排列，排名不分先后）：

总策划：胡文慧、李浩宇、李加赞、李志伟、刘哲理、秦湛、滕义、田英杰、王继刚、王继龙、陶冶、谢攀、徐雷、熊舟、叶晓煜、张云勇、张熙、郑涛

编委会：安常青、董航、范思宇、冯运波、关雪峰、管铭、郝振武、贾国超、贾雅清、蒋武、李超、李大勇、李慧芳、李亚慧、李松鹤、廖林然、梁晨、刘东、卢宇荣、盛杰成、宋立昂、苏俐竹、苏洋、孙世丁、孙学磊、王宝栋、王泰格、王莹、许秀莉、闫小兵、朱育莹、张力天、张景龙、张小梅

一、AI 前沿技术发展

多模态大模型为跨模态理解与生成提供基础能力，智能体通过推理、规划、记忆与工具调用，将模型能力转化为面向目标的任务执行能力；具身智能依托物理载体，在与环境的感知、行动和反馈中实现智能交互。本章重点介绍多模态大模型技术、智能体技术和具身智能技术的发展脉络与主要挑战。

1.1 多模态大模型技术

以 Transformer^[1]架构为基础的大规模预训练推动人工智能从专用模型向通用基础模型演进。随着训练数据、模型规模和计算资源持续扩展，以及指令微调、偏好对齐等技术的发展，以 OpenAI 的 GPT 系列、Anthropic 的 Claude 系列、Google 的 Gemini 系列，以及国内阿里巴巴的 Qwen 系列、智谱的 GLM 系列和深度求索的 DeepSeek 系列等为代表的大语言模型，在语言理解、知识问答、逻辑推理、代码生成和工具调用等方面取得显著进展。然而，纯文本语言模型主要处理离散文本 Token，难以直接感知图像、视频和音频信息。为突破这一限制，研究者通过模态编码、表示对齐和多模态指令微调，将非文本信息映射至语言模型的表示空间，推动大语言模型由文本智能向多模态智能扩展。与此同时，以 Stable Diffusion^[2]为代表的视觉生成模型主要沿扩散模型路线发展，多模态理解与内容生成通常采用不同的架构和训练目标。在此背景下，多模态大模型形成两条相互关联的演进主线：一是扩大模型感知和交互的模态范围，二是推动多模态理解与生成在统一框架内融合。

（1）多模态大语言模型：从文本到视觉，再到全模态融合

多模态大语言模型（Multimodal Large Language Model, MLLM）通常以预训练语言模型为认知核心，通过模态编码器和连接器将图像、视频、音频等非文本信息映射至语言模型的表示空间，形成“模态编码器—连接器—大语言模型”的基本架构。早期模型主要由文本向视觉扩展，如 LLaVA 通过视觉特征对齐和多模态指令微调实现图像理解与视觉对话^[3]。此后，模型逐渐向文本、图像、视频和音频等主要模态的联合处理发展。Qwen2.5-Omni 实现多模态感知和流式文本、语音生成；Qwen3.5-Omni 进一步增强了长上下文、音视频理解和多语言语音交互能力。总体而言，多模态大语言模型正由图文理解向多源信息联合感知、跨模态推理和实时交互演进。

（2）统一多模态基础模型：理解与生成的统一

多模态理解模型通常以自回归语言模型为核心，并以文本作为主要输出，而图像、视频等视觉内容生成常采用扩散模型，两类系统在表示空间、网络结构和训练目标上相对分离^[4,5]。统一多模态基础模型（Unified Multimodal Foundation Model, UMFM）旨在融合感知、推理与内容生成能力，其技术路线可概括为模块化协同和模型级统一：前者由语言模型承担理解、推理与任务规划，再调用扩散模型等外部生成模块，具有较强的灵活性，但存在模块衔接和误差传播问题；后者则在共享主干和训练框架中联合建模语言生成、视觉理解与视觉生成。Show-o 在统一 Transformer 中结合自回归语言建模与离散扩散式视觉生成，可完成视觉问答、图像生成和图像编辑等任务^[6]。这一方向表明，

多模态模型在独立模块功能组合的基础上，正进一步探索模型结构、表示空间和训练目标的深度融合。

（3）研究展望与挑战

多模态大模型仍面临三方面挑战。一是**跨模态表征问题**。文本、图像、音频和视频在信息密度、时间尺度与表示方式上存在显著差异，统一离散化可能导致信息损失或序列膨胀，而分别编码又增加了跨模态语义对齐的难度。二是**数据与训练问题**。高质量多模态数据及真实交互数据相对稀缺，其采集与标注还受到成本、版权和隐私等因素制约；多模态联合训练容易出现模态失衡和能力相互干扰。三是**评估、效率与安全问题**。现有基准难以全面衡量模型的感知、推理、生成、实时交互和跨模态一致性，大规模音视频处理还带来较高的训练与推理成本，并伴随跨模态幻觉、深度伪造和隐私泄露等风险。

1.2 智能体技术

以大语言模型（Large Language Model, LLM）为核心的智能体（LLM Agent）正推动人工智能从响应单次请求向自主规划和持续执行任务发展，并成为人工智能领域最具变革性的研究方向之一。LLM 智能体能够感知环境、理解目标，并自主规划和执行复杂任务^[7]。Wooldridge 和 Jennings^[8]于 1995 年将智能体定义为能够自主感知环境并采取行动以实现目标的实体，早期智能体的知识表示和推理能力有限，许多系统依赖预设规则，难以适应开放环境。以 GPT 系列为代表的预训练语言模型通过大规模预训练获得知识表示与语言处理

能力，为智能体提供了语言理解、常识推理与任务规划基础。思维链提示提升了模型的多步推理能力^[9]，Toolformer 等方法则推动了工具调用能力的发展^[10]，促使 LLM 智能体由研究概念逐步走向工程化与产业应用。

（1）LLM 智能体：早期探索与范式确立

2022 年，思维链（Chain-of-Thought）^[9]通过引导大模型生成中间推理步骤，提升其在多步推理任务中的表现。2023 年，ReAct 框架^[11]将推理与行动交替融合，使模型能够依据工具执行结果和环境反馈调整计划，为智能体的推理与执行闭环提供了框架。同年，思维树（Tree of Thoughts）方法^[12]以树状搜索替代线性推理，使模型具备回溯和自我纠错能力。在记忆与工具层面，Lewis 等^[13]提出检索增强生成（RAG），通过检索外部知识增强生成，为智能体接入“可检索记忆”提供了基础。尽管早期系统的可靠性有限，这些探索仍推动了规划、执行、反馈与迭代机制的发展，为后续智能体系统奠定了基础。

（2）通用 Agent：从工程化探索到规模化应用。

2024 年，LLM 智能体的工程化探索加快。2024 年 10 月，Anthropic 公开测试 Computer Use，使 Claude 可理解屏幕界面并操作鼠标、键盘¹；同年 11 月开源模型上下文协议（Model Context Protocol，MCP），以统一接口连接模型、工具和数据源²。2025 年，智能体走向专业化与标准化。2025 年 2 月，Anthropic 发布 Claude Code，使模型能够修改代码、运行测试并调用命令行工具³；同年 4 月，Google 发布智能体间协议（Agent2Agent，A2A），支持异构

¹ <https://www.anthropic.com/news/developing-computer-use>

² <https://www.anthropic.com/news/model-context-protocol>

³ <https://www.anthropic.com/news/claude-3-7-sonnet>

智能体进行能力发现、信息交换和任务协作，与面向工具及数据连接的 MCP 形成互补⁴。同月，OpenAI 开源发布本地终端编程智能体 Codex CLI⁵。2026 年，发展重点转向长时任务、多智能体协作与跨应用操作。2026 年 3 月发布的 GPT-5.4 进一步增强了长上下文、计算机操作和工具使用能力⁶。

与此同时，国内通用智能体平台也蓬勃发展。2024 年，百度文心智能体平台已支持零代码智能体创建，并通过知识库、插件和工作流编排拓展应用能力；同年 10 月，智谱发布 AutoGLM，使智能体能够感知并操作移动端图形界面，完成连续任务。2025 年 8 月，阿里 Qoder 发布首个版本，依托代码库理解、任务规划与异步执行能力支持复杂开发任务。进入 2026 年，国内产品进一步向个人与专业工作场景延伸。6 月，腾讯发布 WorkBuddy 企业版，连接文档协作、知识沉淀与组织工作流；字节跳动旗下的豆包任务模式可理解网页内容、调用浏览器并完成点击、输入等多步骤操作，同期发布的 TRAE Work 则面向专业用户，覆盖资料研究、内容撰写、数据分析、任务规划与协同工作。2026 年 8 月，DeepSeek 发布开源智能体运行框架 DeepSeek Harness，该框架以“万物皆插件”的开放架构支持智能体组件的灵活组合与扩展⁷。

（3）研究展望与挑战

尽管发展迅速，LLM 智能体从实验原型走向生产部署仍面临三方面挑战。
一是可靠性与安全性不足。链式调用和迭代推理可能使模型幻觉沿任务链累积

⁴ <https://developers.googleblog.com/en/a2a-a-new-era-of-agent-interoperability/>

⁵ <https://learn.chatgpt.com/docs/codex/cli>

⁶ <https://developers.openai.com/api/docs/models/gpt-5.4>

⁷ <https://github.com/deepseek-ai/deepseek-harness>

和放大，在医疗、法律等高风险场景中造成严重后果；自主性增强也扩大了提示注入、工具滥用和多智能体攻击传播等风险，亟须建立覆盖推理、工具、通信及输出审计的纵深防御体系。二是长期记忆能力受限。现有智能体主要通过历史截断、摘要压缩和检索增强管理有限上下文，尚缺乏对不同类型长期记忆的统一组织与动态管理能力，难以实现重要信息保留、过时信息遗忘和新经验持续更新。因此，需要构建分层长期记忆架构，推动记忆从被动存储向主动形成、调用和演化转变。三是评估与标准化体系滞后。静态单轮基准难以衡量多轮策略适应、协作涌现和长期行为漂移，数据污染也会削弱评估可信度；同时，不同平台在工具生态、技能规范和智能体定义上仍存在差异。未来应围绕可靠执行、长期记忆管理和动态评估完善技术能力，推进互操作标准建设，并明确审计与问责机制。

1.3 具身智能技术

具身智能（Embodied AI）旨在使智能体通过物理实体感知、理解并作用于真实世界，是探索通用人工智能（AGI）的研究路径之一。不同于主要在虚拟空间中处理信息的离身智能，具身智能强调认知、身体与环境的闭环交互，使机器人能够在动态环境中自主感知、规划并完成任务。近年来，多模态大模型增强了跨模态理解与交互能力，世界模型提升了环境预测与规划水平，具身智能技术则提供了多样化行动载体，这些技术共同推动具身智能向“感知—推理—规划—行动”一体化方向发展^[14]。

（1）早期探索与技术积累

具身智能的早期研究主要围绕机器人控制、环境感知与强化学习展开，但受感知精度、样本效率和泛化能力限制，难以应对开放环境。进入 2010 年代，深度学习推动了视觉识别、语言理解和运动控制能力的提升，视觉编码器与语言模型分别增强了环境感知和指令理解能力^[14]。与此同时，仿真技术逐渐成为降低训练成本与真实部署风险的重要手段。2024 年推出的 Genesis⁸将多物理场仿真、渲染和机器人训练接口整合于统一平台，为大规模数据生成与策略训练提供了技术支撑。机器人的物理载体也由固定基座机械臂扩展至移动式、四足、人形和仿生机器人，推动具身智能面向更多真实场景发展。

（2）多模态大模型驱动的具身智能

Google 提出的 RT-2^[15]采用视觉—语言—行动（Vision-Language-Action, VLA）范式，将大规模视觉—语言数据中蕴含的知识迁移至机器人控制，使模型能够依据视觉观察和语言指令生成相应动作，提升了机器人在新任务、新物体和新场景中的泛化能力。从系统实现看，具身智能涉及以下相互关联的环节：具身感知负责获取并理解环境信息；具身交互依据语言指令完成目标识别、操作与人机协作；任务规划与执行将高层目标分解为低层动作，并根据反馈动态调整；仿真到现实迁移则通过领域随机化、世界模型和策略适配缩小虚拟训练与真实部署之间的差异。在这些环节中，多模态大模型正由感知或语言处理模块向连接高层认知与低层控制的决策模块拓展。

（3）研究展望与挑战

⁸ <https://github.com/Genesis-Embodied-AI/genesis-world>

具身智能要实现通用、可靠且规模化的部署，仍面临三方面挑战。一是数据与迁移问题。高质量、跨场景、跨设备的真实多模态数据仍较稀缺，仿真与真实环境之间的差异又可能限制训练所得策略的迁移效果。二是规划与推理能力不足。长时任务涉及目标分解、动作衔接、异常恢复和动态重规划，现有多模态模型在常识推理、长期规划及低层操作方面仍存在明显不足。三是评估与安全机制不完善。现有基准对高层语义规划和低层物理控制的联合评价仍不充分；依赖大模型决策的具身智能体还面临后门攻击、指令注入和场景操纵等风险^[16]。未来应重点建设真实多模态数据体系，提升因果推理与长时规划能力，完善覆盖认知决策与物理执行的评估和安全治理机制。

二、AI 安全治理体系

AI 安全治理体系以“安全可靠、公平可信、智能向善”为总体目标，推动人工智能在能力持续演进和应用不断深化的同时，实现风险可识别、过程可管控、行为可干预、责任可追溯^[17,18]。其中，安全可靠强调模型、数据、系统和基础设施在复杂环境下保持稳定运行，能够抵御攻击、故障及异常行为；公平可信关注算法决策、内容生成和服务过程的公正、透明、可解释与可验证，防止偏见、歧视和信息误导；智能向善要求人工智能的研发与应用符合人类共同价值和公共利益，避免技术能力被滥用。围绕这些目标，体系以数据和算力基础设施为支撑，以模型与算法、智能体安全为重点，将安全运营贯穿规划、研发、部署和运行全过程，并由监督与管理提供制度约束和实施保障。

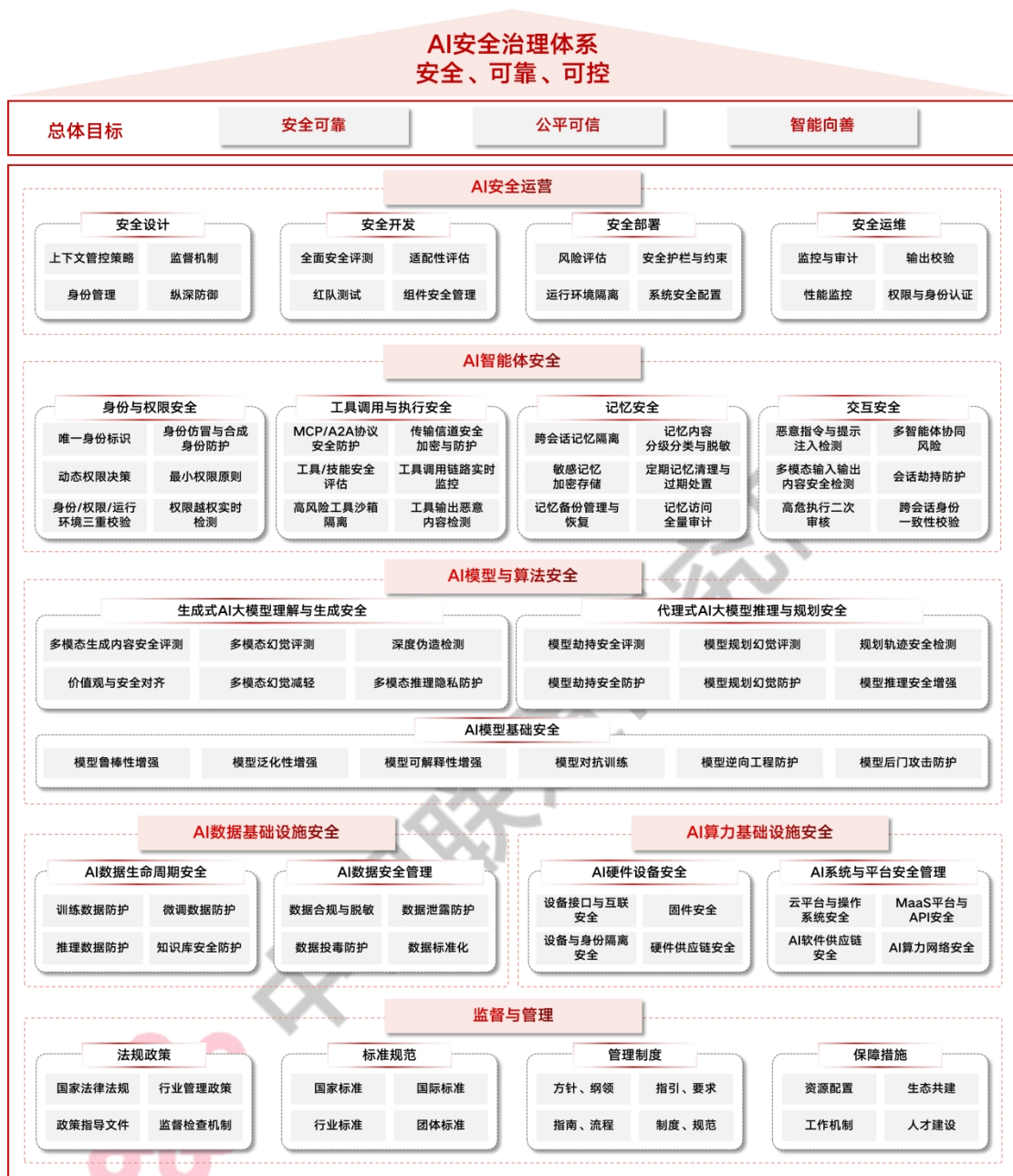


图 2-1 人工智能安全治理体系

2.1 AI 模型与算法安全

AI 模型与算法安全为内容生成、推理规划及上层应用提供能力保障，重点是提高模型输出的可靠性、推理规划的可控性和模型自身的安全性。生成式模型可能出现有害内容、偏见歧视、幻觉、深度伪造及信息泄露，代理式模型则

可能因目标劫持或规划异常引发连续决策风险。这些问题既与训练和模型机理有关，也可能由提示注入、越狱、对抗或后门攻击触发。因此，模型治理需兼顾输出与行为约束、推理规划控制及基础能力增强。

AI 模型与算法安全围绕上述任务形成三方面能力。**一是生成式 AI 大模型理解与生成安全。**开展多模态内容安全和幻觉评测，结合深度伪造检测、价值观与安全对齐、幻觉减轻及推理隐私防护，降低有害或失实内容生成和敏感信息泄露风险。**二是代理式 AI 大模型推理与规划安全。**开展目标劫持、规划幻觉评测与防护，通过推理安全增强和规划轨迹检测识别危险步骤、目标漂移及异常工具依赖，对高风险规划及时阻断或转入人工确认。**三是 AI 模型基础安全。**增强模型鲁棒性、泛化性和可解释性，开展对抗训练，加强逆向工程与后门攻击防护，并通过持续评测和版本验证控制迭代风险。基础安全为理解生成与推理规划提供支撑，三方面能力共同约束模型输出和行为。

2.2 AI 智能体安全

AI 智能体能够感知环境、保持记忆、推理规划并调用工具执行操作，其安全治理重点是控制从任务理解到实际执行的行为边界^[19]。自主性和跨系统连接能力使身份权限滥用、工具调用异常、记忆污染及交互攻击可能沿任务链传导，造成越权操作、信息泄露或服务中断；多智能体协同还可能扩大影响范围。对于具身智能，数字风险可能进一步转化为物理危害。因此，需要将身份、工具、记忆与交互控制贯穿任务全过程，确保关键行为获得授权并可被干预。

围绕任务执行的行为边界，智能体安全包括四方面控制。**一是身份与权限安全。**建立唯一身份标识，防范身份仿冒与合成身份，依据任务和风险动态调整权限，落实最小权限原则，并通过身份、权限和运行环境三重校验及越权检测约束访问。**二是工具调用与执行安全。**加强 MCP/A2A 协议和传输信道防护，开展工具与技能（Skill）安全评估、调用链路监控和工具输出检测，对高风险工具实施沙箱隔离，保留关键执行记录。**三是记忆安全。**实施跨会话隔离、分级分类与脱敏、敏感记忆加密、定期清理及备份恢复，并审计记忆访问，减少污染和泄露。**四是交互安全。**检测恶意指令、提示注入及多模态输入输出内容，防范会话劫持和多智能体协同风险，通过跨会话身份一致性校验与高危执行二次审核保障交互可信。

2.3 AI 数据基础设施安全

AI 数据基础设施为训练、微调、推理和知识增强提供数据支撑，其安全治理重点是保障数据的合规性、机密性、完整性、可用性和可追溯性。来源或授权不清可能引发合规与权益争议，访问控制不足和隐私推断攻击可能导致敏感信息泄露，数据投毒及后门则可能将错误关联传导至模型。由于不同用途的数据在敏感程度、更新方式和使用边界上存在差异，需要将针对各类数据的安全防护与统一管理相结合。

数据安全从生命周期防护和安全管理两方面落实。**一是 AI 数据生命周期安全防护。**训练数据防护重点核验来源、授权、质量与完整性；微调数据防护重

点控制场景数据的敏感性、使用范围及变更；推理数据防护重点减少实时输入、上下文和输出在调用链路中的泄露与滥用；知识库防护重点加强内容准入、更新核验、访问控制和异常检索识别，防范恶意知识注入及越权查询。**二是 AI 数据安全**。建立统一的治理规则、责任分工和流程，将合规与脱敏、泄露防护、投毒防护落实到各类数据处理活动。通过数据标准化统一格式、质量、标识和管理口径，记录来源、授权、访问、流转、变更及处置过程，为生命周期防护提供可执行、可追溯的管理依据。

2.4 AI 算力基础设施安全

AI 算力基础设施由硬件设备、云底座、操作系统、平台软件、接口和算力网络共同构成，任何底层缺陷都可能影响模型训练、推理服务和数据处理的连续性^[20]。硬件、固件和接口缺陷可能造成设备篡改或敏感信息泄露；云平台、操作系统、MaaS 平台及算力网络的漏洞或配置失当，可能引发隔离失效、权限滥用及服务中断。各层相互依赖，风险还可能随节点通信、资源调度和跨域连接扩散，因此需协同保障硬件设备及其承载的系统、平台和网络环境。

算力基础设施安全据此划分为两个相互配合的方面。**一是 AI 硬件设备安全。**对设备接口、外部端口和节点互联实施边界控制与状态核验，规范固件来源、更新和完整性验证，落实设备与身份隔离，防止运维和业务身份混用。将供应链管理贯穿采购、交付、部署、运行、维修和退出环节，保障设备及组件真实、完整、可追溯。**二是 AI 系统与平台安全管理。**加强云平台与操作系统、MaaS

平台与 API、AI 软件供应链及算力网络安全，协同控制账户权限、系统配置、容器镜像、组件依赖、接口调用和跨域资源调度。通过漏洞发现、配置核查、异常监测、资源管控及应急处置，降低运行环境被入侵、篡改或资源耗尽的风险。

2.5 AI 安全运营

AI 安全运营将各模块的安全要求落实到规划、研发、测试、部署和运行全过程，重点是持续发现风险并推动处置与改进。设计边界不清、开发依赖存在缺陷、部署配置不当以及运行中的行为或性能异常，都可能影响系统安全；模型迭代和组件变更还会引入新风险。因此，安全运营需要连接各阶段的评估、验证、监测和处置，使前期控制在后续运行中得到持续检验，并将运行发现的问题反馈至设计与开发。

这一全过程管理通过四个环节衔接落实。**一是安全设计。**结合业务场景和数据敏感程度明确安全边界，建立上下文管控、监督和身份管理机制，采用纵深防御，将应用、模型、数据、工具与基础设施的控制措施相衔接。**二是安全开发。**开展全面安全评测、适配性评估和红队测试，加强组件安全管理，核验第三方依赖、版本变化及能力扩展，减少缺陷向后续环节传递。**三是安全部署。**上线前评估风险，配置安全护栏与约束，落实运行环境隔离和系统安全配置，并保留模型托管、接口开放、权限及变更的审批验证记录。**四是安全运维。**持续监控和审计模型调用、数据访问、工具执行及资源消耗，开展输出校验和性

能监控，强化权限与身份认证，对告警进行定位、分级处置和复盘，形成发现、响应、修复与验证的闭环。

2.6 AI 监督与管理

AI 技术的监督与管理为各技术模块提供制度约束和实施保障，重点是明确治理依据、责任分工和能力要求。不同应用领域在业务目标、数据敏感程度、模型自主程度及风险承受能力上存在差异，统一要求需要与场景实际相衔接。模型、数据、算力、工具及外部服务之间的复杂依赖，也要求治理覆盖跨主体和跨生命周期的责任衔接。为避免要求难以落地、监督与整改脱节，需要同时完善规则、流程和资源保障。

监督与管理通过法规政策、标准规范、管理制度和保障措施协同发挥作用。

一是法规政策。依据国家法律法规明确研发、提供和使用活动的底线要求，通过行业管理政策细化责任与风险控制重点，以政策指导文件明确方向，并通过监督检查核验责任落实和整改情况。

二是标准规范。发挥国家、国际、行业和团体标准的作用，统一技术语言、能力要求和评价尺度，为安全评测、能力建设及监督检查提供依据。

三是管理制度。将 AI 安全方针和要求转化为风险评估、开发测试、上线审批、运行监测及事件处置的具体流程，明确各主体的授权、记录、检查、整改与问责责任，促进各模块协同落实。

四是保障措施。根据业务规模、应用场景和风险水平配置资源，推进生态共建，完善常态化工作机制与人才建设，为技术防护、评测验证、监督检查和应急处置提供持续支撑。

三、AI 安全治理风险分析与挑战

随着 AI 加速融入经济社会关键领域，其风险可能沿技术链条逐层传导并放大，影响系统稳定、数据权益和公共安全，亟须系统防范^[17,21-23]。AI 安全风险自下而上可分为五类：基础设施安全风险，涉及硬件、云平台、MaaS 平台和算力网络；模型与算法安全风险，涵盖鲁棒性、泛化性、可解释性、偏见及攻击威胁；数据安全风险，包括合规、泄露、投毒、隐私和知识产权问题；应用安全风险，涉及多模态、智能体和具身智能；应用场景安全风险则覆盖科研、制造、医疗、金融、政务等重点领域。



图 3-1 AI 安全风险

3.1 AI 模型与算法安全风险

模型安全风险主要包括内生脆弱性与外部攻击威胁^[24-26]。内生风险形成于

模型设计和训练阶段，集中表现为鲁棒性不足、决策过程缺乏可解释性、分布外泛化能力薄弱，以及在性别、地域、种族等敏感维度上的偏见与歧视，其根源涉及训练数据覆盖有限、模型黑箱特性、分布偏移及公平性约束缺失。外部威胁则由恶意主体主动发起，典型形式包括植入隐蔽触发机制的后门攻击、利用微小输入扰动误导模型的对抗攻击，以及通过反复查询接口复制模型能力的模型窃取攻击，可能损害模型的可用性、完整性、机密性与知识产权。

3.1.1 模型内生安全风险

模型内生安全风险是指受模型结构、训练机制、数据分布以及知识表示方式限制，在模型构建和运行过程中存在的安全脆弱性。主要风险包括：

（1）**鲁棒性不足**：由于深度学习模型通常依赖大量训练样本学习统计特征，对输入扰动、异常样本和边界场景的适应能力有限，当实际输入与训练环境存在细微差异时，模型可能产生错误或不稳定输出。

（2）**可解释性不足**：由于深度神经网络具有多层非线性结构和海量参数，模型决策过程通常难以被人类理解和追踪，导致模型输出缺乏透明性，限制了安全审计、异常分析和责任追溯。例如，在医疗辅助诊断场景中，模型可能输出错误诊断结果，但由于缺乏有效解释机制，医生难以判断错误来源并采取修正措施。

（3）**泛化能力不足**：由于训练数据规模、覆盖范围和分布特征有限，模型学习到的知识可能局限于特定场景，当部署环境或输入数据分布发生变化时，

模型性能可能明显下降。例如，模型在实验环境中表现良好，但在真实业务环境中可能出现准确率下降、错误预测或异常行为。

（4）偏见与价值对齐风险：由于训练数据可能包含历史偏见、不均衡样本或不符合预期价值的数据模式，模型可能学习并放大数据中的不公平因素。同时，若模型缺少有效的公平性约束和安全对齐机制，可能生成偏离应用目标、伦理要求或安全规范的输出。

3.1.2 模型外部攻击威胁

模型外部攻击威胁是指攻击者在模型训练、部署和推理过程中，通过数据操纵、输入构造或接口交互等方式对模型安全属性造成的威胁^[27]。典型攻击包括：

（1）后门攻击：由于 AI 模型训练过程依赖大规模数据和第三方组件，攻击者可能通过污染训练数据、篡改模型参数或操纵微调过程，在模型中植入隐藏触发机制。被植入后门的模型通常对正常输入保持预期表现，但在满足特定触发条件时执行攻击者预设行为。例如，攻击者在图像识别模型训练数据中植入带有特定标识的样本，使模型在检测到该标识时输出错误分类结果。

（2）对抗攻击：攻击者通过构造人眼不可察觉或语义变化较小的恶意输入样本，诱导模型产生错误预测。例如，攻击者在自动驾驶环境中对道路标识进行轻微修改，使车辆视觉模型错误识别交通标志并执行错误控制策略。

（3）模型窃取攻击：由于 AI 模型通常通过 API 接口提供服务，攻击者可

以利用大量查询请求分析模型输入输出关系，推断模型功能和参数特征，甚至据此训练替代模型。

3.2 AI 数据安全风险

AI 数据安全风险贯穿数据采集、存储、处理、使用及共享的全生命周期，主要包括治理不足与恶意攻击两类。治理风险源于制度缺失、技术防护薄弱及合规意识不足，具体表现为数据处理违法违规、敏感个人信息泄露、未经授权使用版权内容，以及数据质量低下导致模型性能下降或决策失误。攻击风险则包括通过注入恶意样本或篡改标签实施训练数据投毒，以及利用模型输出开展成员推理、属性推理和数据重建，从而泄露训练集成员身份、个人敏感属性或用户输入历史；此外，数据存储与访问环节还面临非法访问和窃取威胁。

3.2.1 数据安全治理风险

数据安全治理风险贯穿数据采集、存储、处理、使用及共享等全生命周期环节，主要由组织内部管理制度缺失、技术防护薄弱或合规意识不足所引发，主要风险包括：

（1）**数据合规风险**：由于 AI 系统训练和应用需要大量数据支撑，若数据采集、处理、共享过程缺少合法授权和规范管理，可能违反《数据安全法》、《网络安全法》、《个人信息保护法》等相关法律法规要求。该风险主要表现为超范围采集个人信息、未经授权使用数据、数据跨境流动不符合监管要求等。

例如，企业在训练智能客服模型时，未经用户授权直接使用包含个人身份信息的历史交互数据，可能造成数据合规风险。

（2）隐私保护不足风险：由于 AI 模型具备较强的数据关联分析和模式学习能力，如果训练数据脱敏措施不足、数据访问权限控制不严格或隐私保护机制缺失，可能导致敏感个人信息在模型训练、推理或共享过程中被暴露。例如，企业仅对训练数据进行了简单脱敏，模型仍可能通过手机号、住址等信息的组合关联识别出特定个人，从而造成隐私泄露。

（3）知识产权侵权风险：由于 AI 模型训练依赖大量文本、图像、音视频和代码等数据资源，如果数据来源不清晰、授权范围不明确或缺少版权管理机制，可能引发知识产权争议。例如，企业未经授权使用商业代码库训练代码生成模型，可能导致相关权利人的商业权益受损。

（4）数据质量与处理安全风险：由于模型性能高度依赖训练数据质量，若数据存在错误标注、缺失、噪声、过时或逻辑矛盾等问题，可能导致模型学习错误模式，影响模型准确性和决策可靠性。例如，企业使用标注错误的信贷数据训练风控模型，可能导致正常客户被误拒。

3.2.2 数据外部攻击威胁

数据外部攻击可作用于数据存储、访问及模型训练、推理等环节，主要包括：

（1）数据投毒与完整性破坏攻击风险：攻击者通过向训练数据、知识库数

据或运行输入数据中注入恶意内容、修改数据标签、替换关键样本或篡改数据内容，使 AI 系统基于错误数据进行学习或推理。例如，攻击者向大模型训练数据中加入带有恶意偏置的数据样本，使模型在特定任务中产生错误输出。

（2）隐私推断攻击风险：攻击者利用模型输出特征、预测结果或交互反馈信息，通过成员推理、属性推理等方法推断模型训练数据中的敏感信息。例如，攻击者针对医疗 AI 模型开展成员推理攻击，判断某患者医疗记录是否被用于模型训练。

（3）数据重建攻击风险：攻击者通过大量调用模型接口、分析模型响应规律或利用模型记忆特征，尝试恢复训练数据片段、用户输入内容或模型学习到的敏感知识。例如，攻击者通过构造特定查询诱导大语言模型重复生成训练数据中的敏感文本片段。

（4）数据非法访问与窃取攻击风险：攻击者通过攻击数据存储系统、利用权限配置缺陷、窃取访问凭证或滥用 API 接口，直接获取用于模型训练、检索增强或推理的数据资源。例如，攻击者利用云端 AI 平台权限管理漏洞，访问企业私有知识库并窃取内部业务资料。

3.3 AI 应用安全风险

本章围绕多模态理解与生成、智能体应用和具身智能三类场景，梳理 AI 应用中的主要安全风险。重点分析多模态内容、幻觉、对抗攻击与深度伪造风险，智能体感知、记忆、决策、执行及供应链风险，以及具身智能在硬件、感知通

信、自主决策和物理交互中的安全风险。

3.3.1 多模态理解与生成安全风险

（1）多模态内容安全风险

多模态内容安全风险是指 AI 系统在理解、转换、生成和传播文本、图像、音频、视频等信息时，因模型能力缺陷、训练数据偏差、安全对齐不足、审核失效或人为滥用，产生、放大或传播违法有害及不当内容的风险。主要表现包括色情低俗、暴力血腥、极端仇恨、侵害未成年人，以及偏见歧视、隐私与人格权侵害、知识产权侵权和合成内容来源不明等问题。不同模态的安全能力存在差异，单一模态的过滤规则难以充分覆盖跨模态输入输出，模态组合也可能引入新的语义风险，导致内容理解错误或有害内容生成与传播^[28]。

此类风险可能损害个人权益、增加平台审核负担，并引发社会恐慌或群体对立；持续生成的不良内容还可能加重对未成年人的影响。例如，2026 年 2 月，中央网信办通报，部分网络账号利用 AI 生成虚假火灾现场图片、视频，编造和炒作事故，并制作低俗惊悚、血腥暴力视频。网信部门督促网站平台排查整治，依法依规处置账号 13421 个，清理违法违规信息 54.3 万余条⁹。

（2）多模态幻觉

多模态幻觉，是指模型在不存在明确恶意操纵的情况下，生成与输入图像、音频或视频、跨模态上下文以及可验证事实不一致，却在语言或视觉形式上看

⁹ https://www.cac.gov.cn/2026-02/12/c_1772636033171974.htm

似合理的内容。主要表现包括虚构对象，误报属性、数量、位置或关系，错误识别图片文字和语音内容，混淆视频事件顺序，以及错误关联人物、地点、年代或因果知识^[29]。其成因包括训练数据中的图文错配和错误信息、视觉编码压缩造成的细节丢失，以及模型依赖语言先验补全证据缺口。连贯、肯定的输出形式可能掩盖不确定性，后续追问还可能使错误持续累积。

多模态幻觉可能影响高风险场景中的判断与决策。例如，医疗影像中的虚构病灶或异常遗漏可能导致误诊误治；交通和工业场景中的障碍物、仪表读数误判可能触发错误控制；金融与政务场景中的票据、证照或录音误读可能引发拒付、错误审批和责任争议；历史文化场景中的错误配图可能影响公共知识的准确性。幻觉输出若被检索系统、报告生成流程或智能体工具链继续使用，还可能沿应用链条扩散，并重新进入网络数据。

（3）多模态对抗与越狱攻击

多模态对抗与越狱攻击是指攻击者利用图像、文字、音频、视频及其组合，干扰模型判断、指令遵循或安全约束的行为。主要形式包括：通过难以察觉的图像或音频扰动，以及特制贴纸、图案或输入组合，诱导模型分类、定位或回答错误；将恶意指令嵌入图片、截图、二维码、文档或隐藏图层，诱导模型将外部数据作为指令执行；绕过拒答机制和安全约束，诱导模型生成暴力、诈骗、恶意代码等有害内容。不同模态的防护差异，以及 OCR、图像缩放、隐写和跨模态组合处理中的薄弱环节，可能被攻击者利用，以规避单一文本检查。

此类攻击可能导致误识别、安全防护失效、有害内容生成，以及个人信息、凭证和商业秘密泄露。隐藏在图像、音频或文档中的攻击内容较难被用户或仅审查文本的系统识别，还可能随截图、海报和扫描件反复传播。当模型连接邮件、日历、支付或企业知识库等系统时，攻击可能进一步诱发越权发信、修改文件、执行代码或发起交易等操作，使输出风险扩展为实际业务风险。

（4）多模态深度伪造与虚假信息

多模态深度伪造与虚假信息风险是指利用 AI 模型生成、编辑、替换或操纵人物面部、声音、口型、动作、场景及事件叙事，并将其作为真实内容传播，从而误导受众或侵害权益的风险。深度伪造侧重内容合成与操纵技术，虚假信息侧重内容失实及其传播影响。经授权的影视特效、配音和虚拟人并不当然属于虚假信息；风险判断需结合授权、标识、身份冒用及是否用于诈骗、诽谤或操纵舆论等因素。

此类风险可能导致资金损失、账户接管、商业秘密泄露及名誉或人格权益受损；伪造灾情、政策或公共人物言论还可能引发恐慌、挤占应急资源并扰乱市场。深度伪造的传播也可能削弱公众对音视频证据的信任，使相关主体借口“由 AI 生成”否认真正的影像证据，形成所谓“说谎者红利”，增加新闻核验、司法取证和机构沟通成本。例如，2023 年 4 月，福州某科技公司负责人郭先生接到冒充好友的视频通话，对方以外地投标需要 430 万元保证金为由请求资金过账。郭先生未核实款项到账即分两笔转出 430 万元，联系真实好友后发现受

骗。公安机关通报指出诈骗者利用 AI 换脸技术冒充好友¹⁰。

3.3.2 智能体应用安全风险

AI 智能体（AI Agent）是具备自主感知、记忆、推理决策、交互与执行能力的智能系统，是人工智能产品及服务的重要形态。其安全风险涉及感知、记忆、推理决策、交互执行及供应链等环节，并可能沿任务链传导和放大^[30,31]。

（1）感知安全风险：提示注入攻击与感知欺骗

感知阶段接收文本、语音、图像、视频等多模态信息，是外部攻击进入智能体任务链的重要入口，主要风险包括：

直接提示注入：攻击者在用户输入中嵌入恶意指令，试图覆盖系统策略、诱导越权操作或泄露数据。例如，在用户请求中嵌入“忽略之前规则”等指令，诱导智能体偏离原有行为约束。

间接提示注入：恶意指令隐藏于网页、邮件、文档、图像隐写或音频嵌入中，在智能体抓取或解析外部内容时被激活。该类攻击具有较强隐蔽性，可绕过用户主动输入检查，成为智能体应用中的重要攻击路径。

多模态感知欺骗：攻击者利用对抗样本、图像隐写、语音伪造、传感器干扰等方式改变智能体对环境的感知结果，为后续认知与规划阶段提供错误前提。例如，通过修改图像中的关键目标特征，使视觉智能体错误识别对象；通过语音伪造或声纹欺骗诱导语音智能体执行非授权操作。

（2）记忆安全风险：记忆投毒、隐私泄露与级联扩散

¹⁰ https://www.shanwei.gov.cn/gdsw110/zwgk/wgk/content/post_924988.html

该阶段涉及智能体短期上下文（Conversation Context）、长期记忆（Long-term Memory），以及通过检索增强生成（RAG）等机制接入的外部知识库和检索上下文。攻击者可通过污染、窃取或操纵相关记忆内容，使风险从单次交互扩展至长期任务过程，主要风险包括：

记忆投毒：攻击者通过恶意对话、污染知识文档、篡改向量数据库或注入虚假知识等方式，将错误信息或恶意指令写入智能体记忆，使其在后续任务中持续引用受污染内容。例如，在企业 RAG 知识库中植入伪造业务规则，可能导致智能体长期生成错误答案或执行违规操作。

上下文隐私泄露：智能体在多轮交互过程中会积累用户偏好、业务数据、身份信息及历史任务信息，如果缺乏有效的记忆隔离、访问控制和数据脱敏机制，可能导致敏感信息被非授权任务检索或泄露。例如，不同用户或不同业务会话之间的长期记忆共享不当，可能造成跨用户数据暴露。

记忆级联扩散：智能体记忆具有长期保存和跨任务复用特点，一旦恶意信息进入共享记忆库，可能影响多个会话、多个用户甚至多个协同智能体，形成跨任务、跨主体的级联风险。

（3）推理决策安全风险：目标劫持与决策失控

推理决策是智能体任务闭环的决策中枢，负责任务理解、目标分解、路径规划和行动选择。由于智能体依赖大模型进行自主推理，容易受到恶意输入、上下文污染、知识偏差及模型固有缺陷影响，主要风险包括：

模型内生缺陷：大模型存在幻觉、语义理解偏差等问题，可能生成错误判断、错误规划或不符合业务约束的执行方案。在复杂任务场景下，这些模型层缺陷可能在多步推理、规划和自主执行链中被持续放大，造成目标偏移或决策失控。

目标意图劫持风险：攻击者通过提示注入、上下文操纵或任务诱导等方式改变智能体原始目标，使其偏离用户真实意图，执行攻击者设定的非授权任务。例如，邮件处理智能体可能被诱导忽略安全规则，自动执行敏感邮件转发；代码生成智能体可能被诱导生成包含漏洞或恶意逻辑的程序。

（4）执行安全风险：工具滥用、权限失控与多智能体通信风险

执行阶段通过外部系统、API、文件与设备将智能体决策转化为实际操作，部分操作影响较大或难以撤销，主要风险包括：

身份权限失控：智能体调用外部工具、服务或资源时，如果缺乏有效身份认证、权限校验和最小权限控制，可能导致越权访问或非法操作。例如，具备数据库、文件系统或企业业务系统访问权限的智能体可能被诱导执行未授权的数据读取、修改、删除或外发操作。

工具滥用风险：攻击者利用智能体的自主调用能力，将原本用于正常业务功能的工具转化为攻击载体，诱导智能体执行超出授权范围或违背安全策略的操作。例如，攻击者通过部署恶意 MCP 服务端或篡改工具描述信息，诱导智能体调用恶意工具，进而实施上下文窃取、敏感数据外传等攻击。

执行与授权不一致风险：智能体生成的执行动作可能与用户真实意图或授权目标不一致。由于 Agent 执行过程由模型自主规划，可能出现错误理解任务目标、错误选择工具或参数配置错误等情况，导致误操作或业务风险。

多智能体通信与协同风险：在多智能体系统中，身份伪装、消息篡改、虚假任务转交或恶意上下文传播可能影响其他智能体，造成任务偏移、权限扩散和级联错误；若智能体间通信缺乏身份认证、消息完整性校验和最小信任边界，单个智能体失陷还可能进一步影响整个协同链路。

（5）供应链安全风险：生态投毒与级联扩散

智能体高度依赖第三方基座模型、插件、技能、MCP 协议组件、工具服务及外部数据源，供应链链路长、生态耦合度高，主要风险包括：

第三方模型与组件投毒风险：攻击者通过发布或篡改模型、微调模型、插件、技能或软件组件时，在其中植入隐藏逻辑、后门行为或恶意代码，使智能体在特定条件下执行异常操作。

MCP 生态与工具服务供应链风险：随着 MCP 等智能体互操作协议的发展，智能体依赖大量外部 MCP 服务端和工具服务完成能力扩展。恶意或不可信 MCP 服务可能通过伪造工具描述、篡改上下文信息、返回恶意结果等方式影响智能体决策，诱导其调用非授权工具或泄露敏感信息。

3.3.3 具身智能安全风险

具身智能系统通过物理载体与环境持续交互，形成感知、决策、行动与反馈的闭环，典型应用包括机器人、自动驾驶车辆和自主无人机等。其安全风险涉及物理硬件、感知与通信、认知与决策、执行与交互四个层面。

（1）物理硬件安全风险：底层可信破坏与设备失控

物理硬件层是具身智能系统运行的基础，承载计算、感知、控制和执行能力。一旦底层硬件或固件环境被破坏，可能直接影响物理载体及执行机构的安全，并绕过上层软件防护。主要风险包括：

设备可信破坏风险：调试接口暴露、安全启动机制缺失、存储未加密以及物理防拆保护不足等问题，可能导致攻击者获取设备控制权限，篡改固件程序、提取密钥或窃取敏感配置。

硬件供应链植入风险：芯片、模组及底层软件中的漏洞或恶意植入，可能在设备部署后被利用或触发，影响感知、计算和控制能力。例如，恶意硬件模块可能窃取运行数据或干扰设备控制行为。

固件升级安全风险：固件空中升级过程中若缺少完整性校验、来源认证和安全回滚机制，攻击者可能伪造升级包并批量传播恶意固件。例如，攻击者利用升级服务器或传输链路漏洞推送篡改固件，导致受影响的设备批量出现异常行为。

（2）物理环境感知与通信安全风险：环境理解失真与信息通道受控

感知与通信层负责获取视觉、声音、触觉、位置及运动状态等信息，并支持具身智能系统与云端、边缘节点及其他系统或设备之间的数据和指令交互，是系统感知物理环境、获取自身状态及开展协同的重要基础。主要风险包括：

多模态感知欺骗风险：攻击者利用对抗样本、环境伪造、语音欺骗或传感器干扰，误导系统对外部环境和自身状态的判断。例如，特制道路标识可能使自动驾驶车辆误识别交通信息，伪造的定位信号可能使自主无人机偏离预定航线。

传感数据完整性风险：传感数据在采集、传输或处理过程中若遭伪造、篡改，可能使具身系统形成错误状态判断。例如，若激光雷达数据在传输或处理过程中遭到篡改，自动驾驶车辆可能误判障碍物位置并规划不安全的行驶路径。

通信链路攻击风险：系统与云端、控制中心及协同设备之间的数据和指令传输可能遭受窃听、篡改或重放攻击。例如，若无人机的控制链路缺少有效的身份认证和完整性校验，攻击者可能注入伪造指令，改变其飞行任务或路径。

（3）环境建模与自主决策安全风险：认知偏差与行动规划失配

环境建模与自主决策层是具身智能实现从感知到行动的核心环节。该层不仅需要识别任务目标和运行约束，还需要构建物理环境状态表征、预测行动后果并生成安全行为策略；部分系统还会采用世界模型等机制辅助环境预测与规划。主要风险包括：

多模态环境理解偏差风险：具身智能系统根据任务需要融合视觉、声音、触觉、位置等信息，感知模型可能因信息融合能力不足、语义理解偏差或训练数据覆盖不足，产生错误环境认知。例如，自动驾驶系统可能因未能正确关联视觉与测距信息，将危险区域判断为可通行区域，进而引发碰撞或任务失败。

物理常识与因果理解不足风险：大模型具备较强语言推理能力，但对真实物理规律、空间关系和因果关系的理解仍存在局限，可能生成符合语言逻辑但不符合物理规律的决策。例如，机械臂的任务规划模型若高估夹具承载能力，可能生成超出设备能力的抓取方案，造成物品掉落或设备损坏。

安全约束理解不足风险：具身智能系统需要同时满足任务目标、物理规律和安全边界。对安全约束的识别或落实不足，使执行过程偏离安全要求。例如，自主无人机在执行快速巡检任务时，若仅优化航程和速度而忽略剩余电量、气象条件及障碍物间距，可能生成缺乏安全余量的飞行计划。

（4）自主执行与交互安全风险：行为失控与物理危害扩散

自主执行与交互层是具身智能风险作用于物理环境的环节，通过运动控制、作业执行、人机协作和设备联动直接影响现实环境，可能产生不可逆的物理后果。主要风险包括：

设备控制权限滥用风险：攻击者利用应用漏洞、认证缺陷或权限配置错误获取设备控制权限，进而远程操控设备、窃取敏感数据或篡改执行动作。例如，攻击者入侵机器人管理平台，修改机械臂动作参数，使设备执行非法操作。

人机交互安全风险：具身智能系统在与操作人员、乘员、协作人员或周边行人交互时，若身份识别、意图理解、控制权交接或安全距离控制不足，可能造成人员伤害或财产损失。例如，协作机器人若错误判断人员位置，可能在人员进入工作区域时仍保持高速运动，造成碰撞伤害。

3.4 AI 应用场景安全风险

3.4.1 AI 基础电信网络

自智网络是以自动化和智能化为核心，通过持续感知网络状态、智能分析、自主决策与执行，实现网络业务和运维闭环自治的通信网络体系^[32,33]。随着网络规模持续扩大、架构日益复杂、业务需求快速变化，依赖人工处置的传统运维模式已难以满足实时性和精细化要求。自智网络通过融合大模型、大数据、数字孪生与自动化技术，推动网络由依赖人工经验的被动运维，逐步迈向“自配置、自修复、自优化”的高阶自治。近年来，随着模型推理、知识理解和工具调用能力不断增强，AI 在电信网络中的应用已由早期的告警分析、流量预测和辅助决策，拓展至故障预测与修复、无线资源调度、网络规划优化、智能节能、投诉处理和业务保障等场景，并开始形成从状态感知、分析决策到任务执行的闭环流程。相关实践正由局部试点走向生产流程，并逐步覆盖网络规划、建设、维护和优化的全生命周期。当前，全行业正从 L3 向 L4 演进，L4 仍主要处于规划、验证和场景化部署阶段。运营商与设备厂商正在推进智能体在智能运维、故障处置、节能降耗和客户服务等场景中的应用与规模化部署。随着单

点智能应用逐步向跨域协同和端到端自治演进，自智网络有望进一步提升运维效率和资源利用率，缩短故障恢复时间，并成为支撑业务创新、经营增收和绿色低碳发展的重要基础。

尽管自智网络应用不断深化，但迈向 L4 过程中仍面临多重挑战。网络域在设备接口、数据格式和自动化能力方面存在差异，增加跨域协同和系统集成难度；数据质量不足或感知信息缺失，可能造成态势判断偏差，引发故障误判和资源错配。随着生成式 AI 和智能体进入网络决策与执行闭环，模型输出不可靠、智能体协同冲突或自治边界不清，可能诱发错误指令，导致网络拥塞、服务中断或信息泄露。因此，高阶自智网络需提升数据与模型质量，明确人工监督和自治边界，并强化行为可控与过程可审计的能力。

3.4.2 AI 网络安全

随着网络威胁日益复杂和隐蔽，大语言模型的语义理解、代码分析与生成能力为安全自动化提供了新的技术路径。自 2023 年以来，大模型赋能网络安全逐步成为网络安全与软件工程领域的重要研究方向之一^[34,35]。

大语言模型逐步成为网络安全智能化研究的重要技术路线之一，开源模型凭借可定制、可微调等特性在领域适配中发挥重要作用。2025 年以来，前沿大语言模型在网络安全领域的能力出现显著跃迁，并逐步呈现出自主化、长程化和智能体化特征。2026 年 4 月，Anthropic 在公开技术报告中披露，Claude Mythos Preview 能够在主流操作系统、浏览器和开源软件中发现数千个此前未

知的漏洞，其中包括一个存在约 27 年的 OpenBSD 漏洞；在 Linux 内核漏洞研究中，该模型还能够自主发现并组合利用多个漏洞，实现从普通用户权限到更高权限的攻击链构造¹¹。Anthropic 进一步报告，在针对 Firefox 147 JavaScript 引擎漏洞开展的自动化利用开发实验中，此前 Claude Opus 4.6 在数百次尝试中仅成功生成 2 次 JavaScript Shell 利用程序，而 Mythos Preview 在相同类型实验中成功生成 181 次可工作的漏洞利用程序。这表明，前沿模型能力正由“漏洞发现”进一步向“漏洞利用生成”延伸。在此趋势下，安全智能体已成为安全自动化的重要发展方向，其技术架构正由单智能体执行逐步拓展至多智能体分工协作。

总体来看，AI 赋能网络安全具有明显的双重用途和自动化放大风险。一方面，模型幻觉、误报漏报以及日志、威胁情报等外部上下文污染，可能造成错误研判和自动化处置偏差；另一方面，当安全智能体被赋予扫描、漏洞验证、代码执行、凭证使用等高权限工具后，提示注入、目标劫持、权限配置不当或工具误用可能将模型判断错误转化为真实网络操作。随着前沿模型漏洞发现与利用生成能力增强，相关能力还可能降低部分攻击任务的技术门槛；多智能体协作场景下，恶意消息传播和错误决策也可能形成级联影响。因此，应对攻击性能力、工具权限、执行环境和人工审批实施分级控制，并建立完整的行为审计、异常回滚和可追溯机制。

¹¹ <https://www.anthropic.com/glasswing>

3.4.3 AI 编程

代码语言模型（Code LLMs）是人工智能与软件工程交叉领域的重要方向，旨在利用深度学习技术理解和生成源代码。2023 年以来，随着模型规模、推理能力和上下文处理能力持续提升，代码大模型的综合能力显著增强，并逐步由专用代码模型向通用大模型与代码智能体协同发展的方向演进。这些模型在代码生成、理解、调试以及零样本和少样本任务中的适应能力持续提升。与此同时，以 GPT、Claude、DeepSeek、Kimi 等为代表的通用大模型不断增强代码理解、生成、调试和复杂软件工程任务处理能力，推动 AI 编程由单纯的代码生成进一步向代码库级理解和任务执行发展。在此基础上，Claude Code、Codex、Cursor 和 GitHub Copilot 等代码智能体工具与平台快速发展：Claude Code 能够直接在代码库中读取和修改文件、运行命令并完成多步骤开发任务；Codex 能够围绕代码库执行功能开发、缺陷修复、代码审查和重构等端到端软件工程任务；Cursor 将代码智能体深度集成至编辑器和终端环境；GitHub Copilot 也已由早期代码补全扩展至代码库分析、任务规划、代码修改等智能体式开发流程。

尽管 AI 编程技术取得显著进展，当前仍面临多方面安全挑战^[36,37]。一是代码正确性与漏洞引入风险，模型可能生成逻辑错误、不安全 API 调用或带有潜在漏洞的代码；二是幻觉依赖与软件供应链风险，模型可能引用不存在、拼写相近或不可信的软件包和第三方组件；三是间接提示注入风险，仓库文档、代码注释或外部网页中的恶意指令可能影响代码智能体规划与执行；四是高权限

工具执行与越界修改风险，代码智能体能够修改文件、执行命令并调用外部工具，若权限控制、人工确认、沙箱隔离和审计机制不足，规划错误、工具误用或偏离用户意图可能进一步传导至真实开发、构建和 CI/CD 环境。

3.4.4 AI 科学研究

科学研究正在受到大语言模型和智能体技术的持续影响。自 2022 年底 ChatGPT 问世以来，以 GPT、Claude、Gemini 等为代表的大语言模型凭借其在复杂推理、代码生成与工具调用等方面的能力持续提升，逐步渗透至科学研究的多个环节。2025 年以来，该领域的发展主要呈现两条关键脉络：一是以 OpenAI Deep Research¹²、Google Gemini Deep Research¹³等为代表的工具，逐步形成面向文献检索、证据综合和科研辅助等任务的专业化能力，降低了文献调研与知识整合门槛；二是以 AI Co-scientist¹⁴、Agent Laboratory¹⁵等为代表的自主科研智能体不断发展，开始探索从假设生成、实验执行到论文撰写的端到端科研闭环，推动大语言模型由单点科研辅助工具向更高自主性的科研系统演进。

尽管进展显著，大语言模型驱动的科学研究的也面临可靠性、安全性与治理等多重挑战。一方面，模型可能产生事实错误、推理偏差、不准确引用或不可复现的研究结论，其黑箱特性也与科学研究强调的可解释、可追溯和可验证要

¹² <https://openai.com/zh-Hans-CN/index/introducing-deep-research/>

¹³ <https://gemini.google/overview/deep-research/>

¹⁴ <https://deepmind.google/blog/co-scientist-a-multi-agent-ai-partner-to-accelerate-research/>

¹⁵ <https://agentlaboratory.github.io/>

求存在矛盾；另一方面，随着科研智能体逐步具备实验规划、工具调用甚至物理实验执行能力，错误决策、权限滥用以及生物化学等高风险领域的技术滥用可能带来更直接的安全影响。此外，现有科研智能体的评测标准、责任边界和伦理治理机制仍不完善。因此，如何在提升科研自主化水平的同时保障研究过程可信、结果可验证和智能体行为可控，将成为未来 AI 赋能科学研究需要重点解决的问题。

3.4.5 AI 医疗

大语言模型正在推动医疗人工智能由传统的文本理解与辅助分析向通用医学推理和智能化临床协作加速演进。早期研究主要通过医学语料预训练和任务微调提升模型的领域适应能力，在 BERT 架构的基础上^[38]，一系列医学预训练模型在命名实体识别、关系抽取和文本分类等临床自然语言处理任务上取得了较好效果。随着大模型的发展，思维链提示^[9]和检索增强生成（RAG）^[13]等技术进一步增强了模型的医学推理与知识获取能力。Med-PaLM 在美国医师资格考试中达到 67.6% 的准确率^[39]，显示出大模型在专业医学知识与推理任务上的快速进步。随着更强通用模型出现，大语言模型应用范围逐步由医学问答拓展至临床决策支持、多模态诊断和多智能体协作。在临床决策支持方面，MDagents 等框架通过多智能体协作显著提升了诊断准确性^[40]，不同专业角色的智能体能够模拟临床专家进行信息交换、结果分析和协同决策。与此同时，Agent Hospital 等虚拟医疗平台开始探索医生、患者及不同专业智能体之间的持续交互与自主学习，体现医疗 AI 由单模型辅助向多智能体协同发展的趋势。

此外，LLM 还逐步应用于医疗文书、医学教育、模拟培训、药物发现和心理健康支持等场景^[41]。

尽管 LLM 在医疗领域取得显著进展，但距离稳定、可靠的临床应用仍有明显差距。首先，幻觉问题是核心瓶颈，包括与事实矛盾的内在幻觉和编造不存在的引用或信息的外在幻觉^[41]，同时，模型的不确定性估计与校准能力仍不充分，其置信表达与实际正确性之间可能存在偏差，使用户难以准确判断输出可靠性。其次，数据隐私与安全不容忽视，医疗数据的高度敏感性要求严格保护患者隐私，大语言模型可能面临训练数据提取等隐私泄露风险^[42]，智能体应用还可能因上下文共享、外部知识库接入和工具调用扩大敏感信息暴露面。同时，模型的“黑箱”特性与临床决策所要求的可解释、可追溯和可审计原则之间仍存在矛盾^[43]。随着医疗智能体自主决策和协作能力增强，安全评测、行为控制、人工复核、责任界定与监管机制也亟待进一步完善。

四、AI 安全治理政策与标准现状

4.1 AI 安全治理法律法规与政策现状

2025 至 2026 年，全球主要经济体持续推进 AI 监管框架的实施与调整。欧盟、美国和我国基于各自的产业基础、法治体系与治理目标，形成了各有侧重的政策路径，主要进展如下：

（1）欧盟

欧盟AI监管节奏呈现松绑趋势。欧盟以《人工智能法案》（AI Act）¹⁶为人工智能治理的基础性框架，其核心方法是风险分级监管，将AI系统划分为不可接受风险、高风险、有限风险、最小风险四类，实施禁止、严格管控、透明义务、宽松监管的差异化措施，筑牢欧盟AI发展的安全底线与基本权利保障防线。2025年11月19日，欧盟委员会发布《关于简化数字立法框架的提案》（Proposal for a regulation on simplification of the digital legislation）、《关于简化人工智能规则的提案》（Proposal for Regulation on simplification for AI rules）等提案¹⁷，旨在修订一系列现行欧盟数字立法，简化合规流程、延长关键条款过渡期、调整义务分配，平衡监管合规与AI产业创新发展需求。2026年5月19日，欧盟委员会正式发布《高风险AI系统分类指南（最终草案）》（Draft Commission guidelines on the classification of high-risk AI systems）¹⁸，该文件是对欧盟《人工智能法案》第6条的配套解读，为AI系统供应商、部署方及市场监管机构提供统一的高风险判定依据。2026年7月，欧盟委员会发布《人工智能与网络安全行动计划》（EU Action Plan on Cybersecurity and Artificial Intelligence）¹⁹，该计划指出先进AI模型正在改变网络安全格局，需要在利用AI增强网络安全能力的同时，应对AI带来的新型网络风险。

（2）美国

¹⁶ <https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai>

¹⁷ https://commission.europa.eu/law/law-making-process/better-regulation/simplification-implementation-and-enforcement/simplification_en

¹⁸ <https://digital-strategy.ec.europa.eu/en/library/draft-commission-guidelines-classification-high-risk-ai-systems>

¹⁹ <https://digital-strategy.ec.europa.eu/en/library/eu-action-plan-cybersecurity-and-artificial-intelligence>

美国的AI治理战略可概括为松绑促创新。2025年7月23日，美国白宫发布《赢得AI竞赛：美国AI行动计划》（Winning the Race: America's AI Action Plan）²⁰，将加速AI创新、建设AI基础设施、领导国际AI外交与安全作为三大战略方向。2025年12月11日，美国总统特朗普签署《确保人工智能的国家政策框架》（Ensuring a National Policy Framework for Artificial Intelligence）行政令，提出建立联邦统一的AI政策框架，减少监管碎片化、促进AI创新发展。2026年3月20日，美国白宫发布《人工智能政策框架》（National Policy Framework for Artificial Intelligence）²¹及相关立法建议，强调建立全国统一的AI政策框架，促进美国AI创新、提升产业竞争力，减少监管障碍。2026年6月2日，美国总统特朗普签署《推动先进人工智能创新与安全》（Promoting Advanced Artificial Intelligence Innovation and Security）²²行政令，强调以创新驱动人工智能产业发展、避免过度监管，并提出升级现有系统以支持AI驱动的网络安全防御、安全部署前沿模型以及防范AI犯罪行为等要求。

（3）中国

我国统筹推进法律修订、政策制定与标准指引编制，持续完善人工智能治理体系^[44]。在立法层面，2025年10月，十四届全国人大常委会表决通过《关于修改〈中华人民共和国网络安全法〉的决定》，新增第二十条专门规制人工

²⁰ <https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf>

²¹ <https://www.whitehouse.gov/wp-content/uploads/2026/03/03.20.26-National-Policy-Framework-for-Artificial-Intelligence-Legislative-Recommendations.pdf>

²² <https://www.whitehouse.gov/presidential-actions/2026/06/promoting-advanced-artificial-intelligence-innovation-and-security/>

智能领域，明确国家支持人工智能基础理论研究和算法等关键技术研发，推进训练数据资源、算力等基础设施建设，完善人工智能伦理规范，加强风险监测评估和安全监管，促进人工智能应用和健康发展。在政策引导层面，《国务院关于深入实施“人工智能+”行动的意见》将人工智能治理建设作为重要章节，提出完善AI法律法规和伦理准则，提升安全能力水平，完善AI安全监测、风险预警与应急响应体系，防范模型黑箱、幻觉、算法歧视风险。2026年3月，工信部等十部门印发《人工智能科技伦理审查与服务办法（试行）》，面向人工智能科研、技术开发主体，规范AI项目伦理审查流程、审查主体、风险分级处置，推动伦理审查落地。2026年4月，国家网信办等五部门联合发布《人工智能拟人化互动服务管理暂行办法》，自同年7月15日起实施，旨在规范AI拟人化互动服务，保障用户权益，促进安全、健康发展。2026年5月，国家网信办等三部门联合印发《智能体规范应用与创新发展实施意见》，明确坚持安全可控，将智能体安全、可靠、可信作为发展的底线要求，贯穿智能体技术研发、应用部署与推广的全过程，切实防范系统性风险。在伦理与标准层面，2025年9月，国家网信办指导、国家互联网应急中心牵头，联合产学研各方制定《人工智能安全治理框架2.0》^[45]，构建以风险管理为核心的系统性治理思路，将人工智能安全风险划分为技术内生、技术应用、应用衍生三大类型。2026年5月，《人工智能应用伦理安全指引1.0》发布，阐明人工智能应用伦理安全理念与原则，为应用开发、服务提供和应用使用等环节提供安全指引。2026年5月，《人工智能 智能体互联》系列标准发布，该系列标准是我国智

能体互联领域首套国家级标准体系，覆盖智能体互联全流程，旨在实现跨平台、跨系统、跨场景的安全协同。2026 年 7 月，《网络安全标准实践指南——智能体部署使用安全指引》发布，围绕评估、准备、部署、使用和停用五个阶段提出全生命周期安全要求。

整体来看，全球 AI 治理正逐步加强风险控制、产业发展与安全保障之间的协调。欧盟侧重权利保护与风险分级监管，美国侧重创新驱动与战略竞争，我国注重统筹发展与全生命周期风险管控。

4.2 AI 安全治理标准规范现状

ISO/IEC、ITU-T、IEEE、IETF 的相关工作及我国国家标准建设在 AI 安全领域既有交集，也各有侧重。相关标准与研究总体涵盖四个基础方向：一是系统安全治理与风险管理，二是数据与隐私安全，三是模型安全测试与评估，四是生成式人工智能系统、服务和内容安全。ISO/IEC 侧重通用管理、可信性与测试验证；ITU-T 突出数据交互、模型安全、智能体协同与前沿 AI 技术安全；IEEE 关注 AI 可解释性、公平性、隐私计算、模型鲁棒性；IETF 则关注智能体通信、协议安全以及身份与权限管控。我国国家标准强调监管衔接、生成内容治理、基础设施防护和行业应用落地。

(1) ISO/IEC

ISO/IEC 人工智能安全标准化工作主要由 ISO/IEC JTC 1/SC 42 人工智能

分技术委员会（Artificial Intelligence）和 ISO/IEC JTC 1/SC 27 信息安全、网络安全与隐私保护分技术委员会（Information Security, Cybersecurity and Privacy Protection）协同推进。两个分技术委员会已围绕 AI 安全形成 20 余项标准项目，覆盖组织治理、风险管理、功能安全、可信性、鲁棒性、测试验证、伦理与社会影响、网络安全及隐私保护等领域。

具体来看，ISO/IEC 人工智能安全相关标准和项目主要涵盖以下五个方面。

一是在系统安全治理与风险管理方面，ISO/IEC 42001:2023《信息技术 人工智能 管理体系》规定人工智能管理体系要求；ISO/IEC 23894:2023《信息技术 人工智能 风险管理指南》提供风险管理方法；ISO/IEC CD 27090《网络安全 人工智能 人工智能系统安全威胁与危害应对指南》提供面向 AI 系统安全威胁与危害应对指南；ISO/IEC 22440 系列《人工智能 功能安全与人工智能系统》分别从风险、要求、指南和应用示例等层面规范 AI 功能安全。**二是在数据与隐私安全方面**，ISO/IEC TR 27563:2023《人工智能使用场景中的安全与隐私 最佳实践》构建安全与隐私实践框架；ISO/IEC FDIS 27091《网络安全与隐私 人工智能 隐私保护》聚焦 AI 隐私保护。**三是在模型安全测试与评估方面**，ISO/IEC 24029 系列《人工智能 神经网络鲁棒性评估》聚焦神经网络鲁棒性评估；ISO/IEC TS 42119-2:2025《人工智能测试 第 2 部分：人工智能系统测试概述》和 ISO/IEC AWI TS 42119-7《人工智能测试 第 7 部分：红队测试》将安全要求转化为标准化测试活动。**四是在生成式人工智能安全方面**，ISO/IEC CD TS 25568《信息技术 人工智能 生成式人工智能系统风险应对指南》分析

风险及控制措施；ISO/IEC AWI 25590《信息技术 人工智能 生成式人工智能应用输出数据质量指南》关注生成式 AI 输出内容质量。此外 ISO/IEC 还关注 **AI 可信性、伦理与透明度** 方面，例如 ISO/IEC TR 24028:2020《信息技术 人工智能 人工智能可信性概述》梳理可信要素；ISO/IEC TR 24368:2022《信息技术 人工智能 伦理与社会关切概述》关注伦理和社会影响；ISO/IEC 12792:2025《信息技术 人工智能 人工智能系统透明度分类》建立 AI 系统透明度信息分类。

总体来看，ISO/IEC 人工智能安全标准注重通用性、过程完整性和方法可验证性，在管理体系、风险表述、可信属性、隐私保护、功能安全和测试方法之间建立衔接。其标准化工作正由框架性指导向技术要求、评估方法和测试流程延伸，体现覆盖全生命周期、管理与技术并重的治理思路。

(2) ITU-T

ITU-T 人工智能安全标准化工作主要由 ITU-T SG17 下设的 Q7、Q8 和 Q16 等研究组推进，其中 Q16 是专门面向 AI 安全领域的研究组。ITU-T 已围绕通用 AI 系统、生成式 AI、基础模型、智能体、具身智能等领域布局 30 余项 AI 安全标准和项目。

具体来看，ITU-T 人工智能安全相关标准和项目主要涵盖以下五个方面。
一是在系统安全治理与风险管理方面，ITU-T X.2211《人工智能系统安全要求》关注系统安全要求；ITU-T X.2213《生成式人工智能应用服务安全指南》聚焦

生成式人工智能服务安全；X.sr-Alapi《人工智能服务 API 安全要求》关注 AI 服务安全。二是在数据与隐私安全方面，X.sg-GenAId《生成式人工智能数据处理安全指南》关注生成式 AI 数据处理安全；X.SG-AIAD《人工智能智能体数据安全指南》则把数据保护延伸至智能体运行环境，关注智能体在感知、规划、执行、记忆等环节面临的数据安全风险及防护措施。三是在模型安全测试与评估方面，XSTR.AI-GSB《基础模型安全基准指南》聚焦基础模型安全评估基准；XSTR.sem-AIA《人工智能智能体安全评估方法》聚焦 AI 智能体安全评估方法。四是在生成式人工智能安全方面，X.GenAI-FT《生成式人工智能模型微调安全指南》、X.AI-gcd《人工智能生成内容检测指南》关注生成式 AI 模型微调与内容生成；X.sgMMFM《多模态基础模型安全指南》聚焦多模态基础模型安全风险及应对措施。五是在智能体与前沿 AI 技术安全方面，X.aaia-sec《自主人工智能智能体安全框架与技术要求》聚焦智能体安全框架；X.S-AIAI《人工智能智能体调用安全框架与要求》关注智能体调用安全；X.sg-AIAof《云侧人工智能智能体编排框架安全指南》聚焦云侧智能体安全编排；X.sg-eAI《具身智能系统安全要求与指南》聚焦具身智能系统安全。

总体来看，ITU-T 人工智能安全标准布局正由通用安全要求向模型、数据、工具调用、服务接口及应用场景延伸，并持续拓展至生成式人工智能、多模态基础模型、智能体和具身智能等前沿领域。

(3) IEEE

IEEE 人工智能安全标准化工作主要由 IEEE 计算机学会人工智能标准委员会（IEEE Computer Society/Artificial Intelligence Standards Committee (C/AISC)）推进。IEEE 已围绕人工智能安全形成 20 余项标准和项目，覆盖可解释性、公平性、隐私计算、鲁棒性测试、大模型评估、生成式人工智能、智能体等领域。

具体来看，IEEE 人工智能安全相关标准和项目主要涵盖以下六个方面。一是在系统安全治理与风险管理方面，IEEE P3395《AI 模型保障、控制与防护技术实施标准》面向恶意使用、组织风险和失控风险提出模型保障与风险缓解策略；IEEE P3396《AI 风险、安全、可信与责任定义及评估推荐实践》构建涵盖 AI 风险、安全、可信与责任的定义及评估框架。二是在数据与隐私安全方面，IEEE P3403《大语言模型训练数据处理推荐实践》规范大语言模型训练数据处理流程和评价方法。三是在模型安全测试与评估方面，IEEE 3129-2023《基于 AI 的图像识别服务鲁棒性测试与评估标准》和 IEEE 3168-2024《采用机器学习的自然语言处理服务鲁棒性评估测试方法标准》分别面向图像识别、自然语言处理服务的鲁棒性评估；IEEE 3378-2025《大规模深度学习模型评估框架与流程标准》规定大规模模型的任务、指标和要求；IEEE P3157《计算机视觉机器学习模型脆弱性测试推荐实践》关注模型脆弱性。四是在生成式人工智能安全方面，IEEE P7018《生成式预训练 AI 模型安全与可信要求标准》关注模型安全风险、隐私泄露等问题；IEEE P3419《大语言模型评估标准》聚焦大语言模型的通用性、智能性、效率与安全评估；IEEE P7022《企业应用中可信生成式与

智能体 AI 要求及评价准则标准》面向幻觉、引用错误和工具误用等失效情形提出评价指标与持续监测要求。五是在可解释性、公平性与可信评价方面，IEEE 2894-2024《可解释人工智能架构框架指南》明确可解释人工智能的定义、方法类型、架构蓝图及性能评价要求；IEEE P2976《可解释人工智能系统要求与分类标准》进一步规定可解释性要求和等级分类。六是在 AI 应用安全方面，IEEE P3481《网络安全专用大语言模型功能要求标准》关注网络安全大模型技术。

总体来看，IEEE 人工智能安全标准以技术方法、量化指标、测试流程和工程可实施性为主要特征，在可解释性、公平性、隐私计算、鲁棒性和大模型评估方面提出较为细化的技术要求。其标准布局正由单项能力和特定任务评估向生成式人工智能和智能体安全，以及全生命周期风险与责任治理拓展。

（4）IETF

互联网工程任务组（IETF）负责制定互联网核心协议和技术规范，相关工作涵盖互联网架构、应用协议、网络安全等领域。IETF 依托工作组、研究组和研讨会等开放机制，围绕 AI 智能体通信与发现、安全协作、身份与权限控制以及内容使用偏好等需求，开展互联网协议层面的标准化探索和前瞻性研究。

目前，IETF 主要围绕以下四个方面开展与 AI 安全相关的协议与标准研究。

一是在智能体身份与权限方面，IETF 依托 Web 授权协议工作组（OAuth）²³和

²³ <https://datatracker.ietf.org/wg/oauth/about/>

多系统环境工作负载身份工作组（WIMSE）²⁴开展相关研究。其中，OAuth 工作组主要研究用户如何授权第三方应用访问资源，相关机制可用于限定智能体的数据访问范围、工具调用权限及其代表用户执行任务的范围，避免智能体获得超出任务需要的权限；WIMSE 工作组主要研究智能体、工具和后台服务等软件工作负载的身份验证及跨系统安全调用机制，以降低身份冒用和未经授权访问的风险。二是在面向 AI 训练与推理的高性能网络方面，IETF 建立快速网络通知工作组（FANN）²⁵，主要研究链路故障、性能劣化和拥塞等网络异常的快速通知机制。三是在 AI 辅助网络运营方面，网络管理运营工作组（NMOP）²⁶正在讨论 AI 和智能体辅助网络运营的相关提案；IETF 网络管理研究组（NMRG）²⁷正研究 AI 与网络管理结合、网络数字孪生等议题，相关工作多处于研究或草案阶段。四是在智能体通信与安全协作方面，IETF 围绕智能体通信安全、网关协作安全等议题，组织了智能体通信协议研讨会（AgentProto）²⁸、智能体发现与工作负载研讨会（DAWN）²⁹以及动态多智能体安全协作研讨会（DMSC）³⁰。其中，AgentProto 主要研究不同企业的智能体之间，以及智能体与工具之间的通用通信方式，以支持消息交换、任务传递与协同执行；DAWN 主要研究网络中智能体、工具和服务的发现机制，帮助调用方了解其位置、能力与适用性；DMSC 主要研究多个智能体通过网关协作时的请求安全转发、任务协调与

²⁴ <https://datatracker.ietf.org/wg/wimse/about/>

²⁵ <https://datatracker.ietf.org/doc/charter-ietf-fann/>

²⁶ <https://datatracker.ietf.org/group/nmop/about/>

²⁷ <https://datatracker.ietf.org/rg/nmrg/about/>

²⁸ <https://datatracker.ietf.org/wg/agentproto/about/>

²⁹ <https://datatracker.ietf.org/doc/bofreq-farrel-discovery-of-agents-workloads-and-named-entities/>

³⁰ <https://datatracker.ietf.org/doc/bofreq-wang-dynamic-multi-agent-secured-collaboration-dmsc/>

访问策略执行机制。

总体来看，IETF 正围绕生成式 AI 和智能体应用带来的网络协议需求，逐步拓展相关标准化研究，内容涵盖智能体通信与发现、安全协作、身份与权限控制、AI 训练与推理网络、AI 辅助网络运营等方面。当前，相关 AI 议题的多项工作仍处于专题研讨、工作组讨论或草案制定阶段，相关标准体系尚在形成。随着智能体互联网相关研究不断推进，智能体身份认证、安全交互和跨系统协议互通等机制将成为进一步研究和完善的重点。

（5）国家标准

我国人工智能安全标准化工作统筹推进顶层设计、标准体系建设、治理框架完善和组织机制建设。2024 年 6 月，工业和信息化部、中央网信办、国家发展改革委、国家标准委联合印发《国家人工智能产业综合标准化体系建设指南（2024 版）》，对人工智能标准化工作作出系统谋划，强化全产业链标准协同，为人工智能产业高质量发展提供技术支撑。2025 年 1 月，全国网络安全标准化技术委员会秘书处组织编制《人工智能安全标准体系（V1.0）》（征求意见稿）并公开征求意见，进一步明确人工智能安全标准体系建设方向。2025 年 9 月，《人工智能安全治理框架》2.0 版正式发布，在 1.0 版基础上结合技术发展和应用实践，持续跟踪风险变化，优化风险分类、探索风险分级，并动态完善防范治理措施。2026 年 3 月，全国网络安全标准化技术委员会组建人工智能安全标准工作组（WG9），主要开展人工智能安全标准现状与趋势调研、标准体系研

究及相关标准研制，进一步健全人工智能安全标准化组织机制。在上述政策和组织机制推动下，我国围绕人工智能安全形成 30 余项标准和项目，内容覆盖数据、算法、模型、计算平台、服务和行业应用，突出监管衔接、内容治理、责任落实、基础设施防护和场景化实施。

具体来看，我国人工智能安全国家标准及相关工作主要涵盖以下六个方面。

一是在系统安全治理与风险管理方面，GB/T 45081—2024《人工智能 管理体系》为组织建立管理体系提供框架；GB/T 45654—2025《网络安全技术 生成式人工智能服务安全基本要求》将安全责任要求落实到语料、模型、服务和用户管理；《网络安全技术 人工智能安全能力成熟度评估方法》用于评价组织安全能力。二是在数据与隐私安全方面，GB/T 45652—2025《网络安全技术 生成式人工智能预训练和优化训练数据安全规范》规范数据来源、处理、质量和安全管理；GB/T 45674—2025《网络安全技术 生成式人工智能数据标注安全规范》覆盖标注平台、规则、人员和核验要求。三是在模型安全测试与评估方面，GB/T 42888—2023《信息安全技术 机器学习算法安全评估规范》规定机器学习全生命周期的安全评估要求；《网络安全技术 人工智能基础模型安全测试方法》和《网络安全技术 智算云服务安全评估方法》将评价对象拓展至基础模型和算力服务。四是在生成式人工智能安全方面，GB 45438—2025《网络安全技术 人工智能生成合成内容标识方法》规定生成合成内容的标识要求；《网络安全技术 人工智能生成内容检测技术实现指南》《网络安全技术 人工智能深度合成安全规范》和《网络安全技术 人工智能代码生成服务安全要求》

分别关注内容检测、深度合成和代码生成中的安全风险。同时，《智能体应用安全基本要求》是我国面向智能体应用的强制性安全国家标准，划定智能体安全准入底线，防范自主决策带来的安全风险，引导智能体安全可信落地应用。

五是在 AI 计算基础设施与开发环境安全方面，GB/T 45958—2025《网络安全技术 人工智能计算平台安全框架》和《网络安全技术 人工智能训练及推理框架安全要求》聚焦 AI 计算平台、开发框架安全。六是在行业应用安全方面，相继发布了教育、金融、应急管理、卫生健康、能源、广播电视、政务等垂直领域指导性技术文件编制计划，体现按行业风险配置安全要求的特点。

总体来看，我国人工智能安全国家标准以风险可控、责任可落实和要求可检查为主线，兼顾基础能力建设与重点场景治理。其标准布局正由原则性治理要求向数据、算法、模型、平台、服务和行业应用等环节的具体要求细化，并持续完善生成内容治理、基础设施防护、智能体安全和行业场景规范，呈现监管衔接更加紧密、技术要求更加可执行、应用覆盖更加广泛的发展趋势。

五、AI 安全治理技术解决方案与案例

5.1 中国联通人工智能安全治理一体化解决方案

中国联通面向人工智能安全前沿技术攻关与产业应用需求，围绕生成式人工智能与代理式人工智能两大核心场景，打造人工智能安全治理一体化解决方案，构建涵盖“有害内容防范”和“有害操作管控”双维度的全栈安全能力体系。方案向下覆盖大模型内生安全、数据安全与基础底座安全，向上延伸至智

能体身份权限、工具调用与全生命周期运营治理，打造百万级高质量评测数据集、安全评测智能体、可解释 AI 安全裁判大模型等核心技术资产，开展价值观对齐、幻觉治理、神经元激活分析等关键技术攻关。依托技术积累，中国联通在 ITU-T 牵头制定 10 余项 AI 安全领域国际标准，累计产出 30 余项发明专利，并在相关领域重要学术会议和期刊发表多项研究成果，形成 20 项以上 AI + 垂直行业落地应用，并持续联合产业伙伴共建 AI 安全生态，推动我国人工智能产业朝着有益、安全、公平的方向健康发展。



图 5-1 中国联通 AI 安全治理一体化解决方案

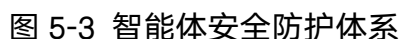
面向生成式人工智能场景，中国联通构建集多模态大模型安全评测与可信增强于一体的技术体系，同步布局可解释人工智能核心技术。在安全评测方面，构建面向多模态理解与生成模型的体系化攻击技术矩阵，涵盖角色扮演、思维链劫持、目标劫持等生成侧攻击手段，以及显式文本嵌入、语义双关注入、跨模态拼接等理解侧攻击方法；同时构建涵盖事实性幻觉、属性错误、关系幻觉

等内容多维度幻觉评测题库，结合通识与垂直行业知识，实现对大模型安全能力的系统化验证。在可信增强方面，以大模型安全对齐数据构建和跨模态安全对齐策略为核心，打造内生安全增强技术体系，并从输入侧、推理侧、输出侧全链路部署幻觉减轻技术，降低模型生成幻觉内容的风险。在此基础上，开展可解释 AI 技术，通过可解释安全裁判大模型与神经元激活分析可视化工具，提升大模型决策过程的可解释性与可审查性，降低模型偏见与错误决策可能带来的社会与经济风险。



图 5-2 中国联通生成式 AI 安全技术全景图

针对代理式人工智能在自主规划、工具调用和多智能体协作中面临的新风险，中国联通构建多位一体的智能体安全防护体系。该体系以分层纵深防御为核心设计原则，贯通技术防护与运营治理两大体系，覆盖智能体从基础模型、边界交互、工具执行到身份权限、记忆数据、基础底座、运营治理的完整技术链路，贯穿智能体创建、运行、迭代到销毁的全生命周期，强化对各环节安全



5.2 百度智能体安全解决方案

百度面向企业智能体规模化应用提出 Agent SOC（安全运营中心）解决方案，以“看得清、管得住、敢放权”为目标，从四个方面构建协同治理体系：

一是构建统一安全中枢，将资产管理、风险感知、策略治理和响应处置贯通起来，并围绕思考、行动和反馈设置控制点。二是部署 Agent 安全护栏，由在线阻断、指令防护和轨迹审计等模型协同驱动，结合内容、意图、上下文、工具调用与行为进行实时研判，在关键节点实施拦截、降权或审计。三是治理第三方 Skill 和插件，建立“来源接入—安全检测—权限声明—上架审核—运行监控—风险处置”的六段式信任链。四是推进安全运营与全生命周期治理，通过统一台账管理智能体、插件和运行环境。



图 5-4 百度 Agent SOC 总体架构

百度安全护栏可以在保持检出能力的同时控制误伤：在线阻断模型召回率为 96.4%、误伤率为 0.09%，指令防护模型召回率为 94.9%、误召率为 0.45%，轨迹审计模型召回率为 96.62%、误伤率为 0.54%。这些指标表明，方案能够兼顾业务连续性与高风险操作拦截，识别语义攻击和异常工具调用，同时避免过度阻断正常任务。统一态势、攻击链研判与自动化处置有助于减少工具间的信息割裂。在供应链治理方面，百度已覆盖 Skill 主流仓库，源同步达到分钟级，

月均扫描超过 100 万次，累计拦截风险超过 1 万件，月均审核 4.8 万件，并处置高危风险 82 个；能力覆盖 DuClaw、DuMate、RedClaw、OpenClaw、Hermes 等平台。Agent SOC 已在百度内部使用场景和业务系统中实现 7×24 小时运行，覆盖 DuMate、DuClaw、RedClaw 及 Agent Infra 等产品。由此形成的实际价值，是将资产发现、风险评估、实时防护、审计溯源和供应链治理连接为统一闭环，为智能体扩大权限、接入生产流程提供可验证、可回收的安全边界。



图 5-5 百度 Agent 安全护栏

5.3 基于中兴 AIOS 的智能体安全治理方案

中兴 AIOS 是中兴通讯面向 AI 时代推出的智能化产品的技术底座，支撑电信网络、企业、家庭、个人智能体应用和多智能体协同。基于中兴 AIOS 的智能体安全治理方案以“安全内生、全程可信”为原则，将传统网络安全与 AI 原生安全结合，覆盖设计、开发、训练、部署、运行和迭代全过程，形成四层治理体系：一是夯实基础安全与身份信任。基础层安全继承身份管理、访问控制、

TLS/SSL 加密、主机加固和行为监控等能力，为智能体划定运行的安全边界；AI 原生层安全面向感知、记忆、推理、规划和执行中风险，为每个智能体建立唯一、可验证的数字身份，并与代理用户、角色、属性和授权来源绑定，为动态授权、细粒度访问和审计追溯提供统一信任基础。二是实施动态最小权限。方案依据业务目标、上下文和风险状态生成最小权限集合，仅授予任务必需的工具调用权、数据访问权或控制指令权；一旦发现意图偏移、越权或异常外发时，将触发告警、暂停、降权、隔离或人工确认。三是推动研发安全左移和安全内生。通过模型安全微调、恶意指令训练、架构安全设计、安全提示词、代码与依赖扫描以及红蓝对抗、模糊测试等措施前置防护，并将插件、Skill 漏洞和许可证检查接入 DevSecOps 及 CI/CD 流程。四是构建运行防护与安全运营闭环。方案涵盖服务入口、工具服务、智能体工作流、安全监测和审计五大环节，通过安全网关、可信技能广场、独立执行沙箱、大模型护栏、智能体护栏等提供多层次的防护，并建立统一安全管理中心，实现日志采集、基线分析、风险告警与追溯审计。通过上述治理方案，基本构建了“事前预防、事中管控、事后追溯、应急响应、生态共治”的智能体防御体系。

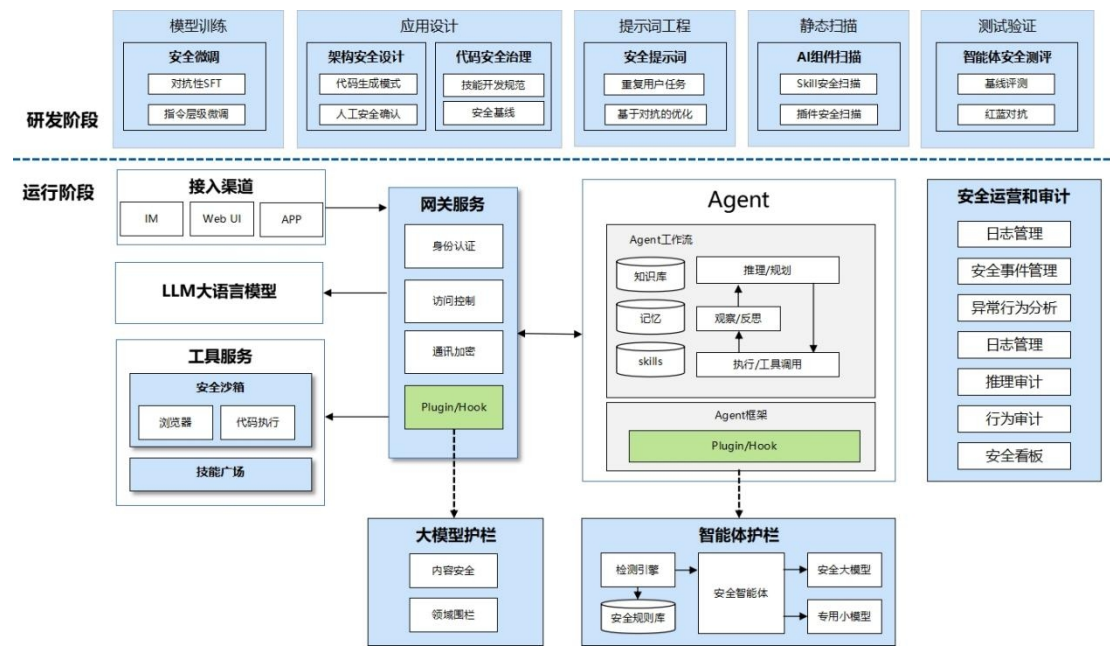
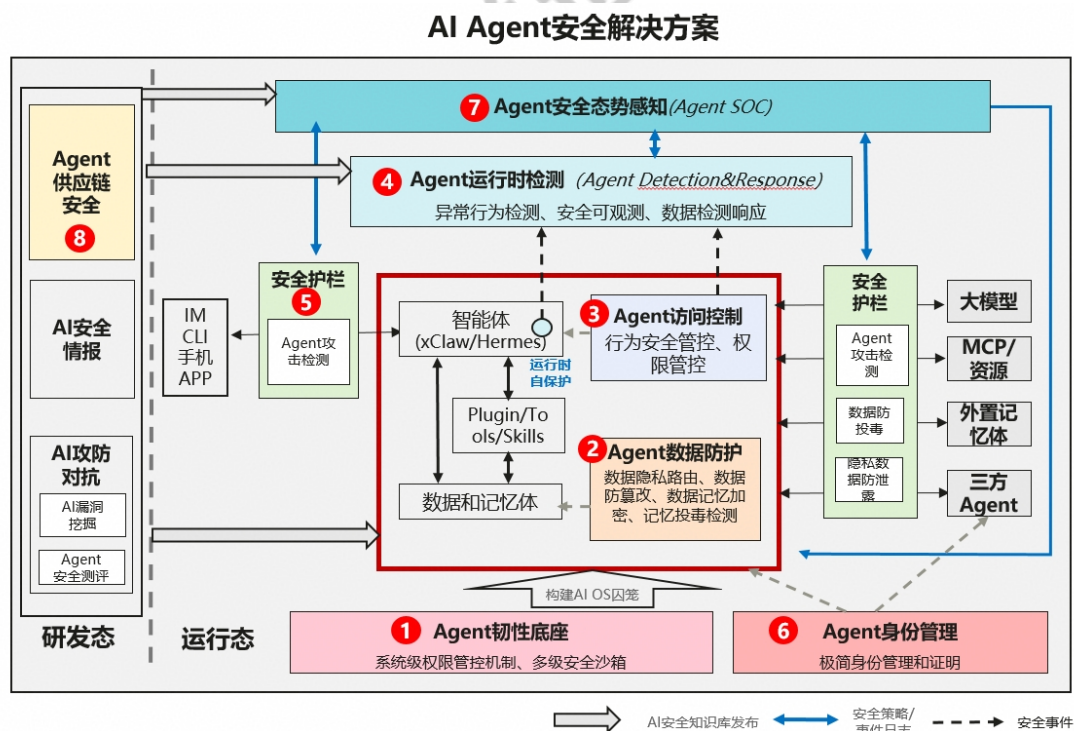


图 5-6 中兴 AIOS 智能体安全治理方案

5.4 华为智能体安全治理方案

华为 AI Agent 安全解决方案以 Agent 韧性底座、数据防护、访问控制、运行时检测、安全护栏、身份管理、安全态势和供应链安全八类能力，构建“事前准入、事中拦截、事后审计”的纵深防御闭环：一是强化供应链准入和身份确权。供应链安全建立模型、代码、插件和 MCP 服务的 AI 物料清单，结合完整性校验、静态与动态扫描、来源签名和漏洞情报控制引入风险；身份管理为非人类身份和多智能体协作建立数字证书、唯一指纹、双向认证及生命周期管理，防范影子智能体、身份仿冒和匿名调用。二是实施事中多层控制。数据防护集中管理 API、数据库和 SSH 凭据，清洗短期上下文与长期记忆，并对 RAG 检索实施权限对齐、敏感数据识别和语义去毒；访问控制遵循最小特权和最小

代理权，对工具、API 及 MCP 通道实施 RBAC、ABAC、动态授权和高风险人工确认；安全护栏在输入输出和知识上下文层识别提示词注入、越狱、内容违规、隐私泄漏与数据投毒，运行时检测汇聚主机、网络、推理链和工具调用日志，通过行为基线发现异常频率、批量提取、横向移动和失控执行；当模型或工具被操纵时，韧性底座利用轻量化安全沙箱、可信执行环境、系统调用白名单、文件目录限制和计算资源配额，将破坏性代码限制在隔离环境内。三是完善安全态势与持续审计。Agent 安全态势统一发现资产和暴露面，构建 AI-BOM、风险画像、攻击路径和可达性关系，对照安全基线与合规要求持续评估，并关联运行时告警、身份、权限、数据和供应链信息，支持从单点事件追踪到完整攻击链复盘。八项能力由此通过统一策略与审计信息形成端到端协同。



从企业智能体治理的应用需求看，供应链准入与身份确权首先回答“组件是否可信、动作由谁发起”，为模型、插件、MCP 服务和非人类身份建立清晰的责任边界；数据防护和访问控制将凭据、知识、记忆、工具及权限纳入任务上下文，压缩智能体长期持有高权限带来的风险；安全护栏与运行时检测分别从语义和行为层识别攻击，使提示词诱导、越权调用和多步异常具备可观测、可拦截、可追溯的处置条件。在此基础上，韧性底座将模型推理的不确定性转化为基础设施层的确定性隔离，通过沙箱、白名单和资源配额限制失控影响；安全态势则把分散的资产、身份、组件、权限和告警汇总为风险暴露面与攻击路径，为安全团队提供统一审计依据。整体方案以身份、权限、数据、语义、行为、供应链和基础设施多层防线形成止损机制，使企业在赋予智能体生产权限时，仍能保持最小授权、人工复核、异常隔离和事后追责能力。

5.5 360 全模态大模型安全卫士 R3.0 解决方案

360 全模态大模型安全卫士 R3.0 面向大模型和智能体应用全生命周期，以“以模治模、以测促防”为核心思路，形成四层防护体系：一是建设统一平台并实现低侵入部署，将模型资产治理、系统风险检测、模型安全评测和智能体应用防护集中管理，采用旁路、反向代理等方式接入，支持 SaaS 服务和私有化部署；规则引擎与 AI 检测引擎协同监控用户输入、模型推理、知识检索、工具调用和结果返回，通过策略编排、告警响应、合规报表和 API 安全网关实现集中运营。二是开展全模态输入输出检测。入口侧检测文本风险、图像隐写、

音频隐藏通道和跨模态信息夹带，对图像、音频执行 OCR 或 ASR 后进行文本级复核，并通过多模态联合意图识别判断组合输入的真实目的；推理侧关注思维链和异常语义行为；输出侧审核文本、图片、音频、视频等内容，识别敏感信息、违法有害信息和 AIGC 内容。三是防范跨模态越狱、深度伪造并强化内容溯源，综合运用跨模态语义一致性、频域和生物信号特征、隐形数字水印及存证机制，降低单一模态检测被绕过的风险。四是覆盖提示词、数据、模型和 Agent 安全全生命周期，可识别越狱、注入和恶意指令，审核训练数据与隐私合规，评估对抗样本、模型窃取和后门风险，审计工具调用、权限与智能体行为；产品支持 100 余种攻击模式，并提供多模态文件审核、图片策略运营、离线直评、自定义拒答库、离线内容包升级及上下文历史违规识别等功能，由此形成事前评测、事中防护和事后调优闭环。



图 5-8 360 大模型安全卫士总体架构



图 5-9 360 从模型到智能体的安全防护体系

5.6 奇安信智能体安全解决方案

奇安信以“大模型是思想、智能体是行动，AI 安全核心是管住智能体的手脚”为出发点，在不大幅改造现有业务架构、不改变智能体应用流程的前提下，为模型、工具、访问、行为意图和结果五大安全域叠加 AI 安全层，按“双闸门打底、行为管控盯防、终端识别补漏、权限行为升级”四步形成四层能力体系：

一是构建控制平面与数据平面分离的总体架构。控制平面由智能体安全管理平台统一承担资产发现、全量审计、策略管控、威胁检测、身份权限编排、攻击溯源和合规报告；数据平面由流量侧和端侧探针组成，可独立部署、按需组合、随智能体规模扩容，支撑智能体统一云化、独立网段隔离、网关前置的部署基线，避免先上线再返工。

二是形成流量与端侧协同防护。流量侧布设“双闸门”：大模型卫士守在模型出入口管内容合规，接住国家监管要求、实现合规免责；AI 安全网关观察 Prompt、工具调用和上下文，管控提示词注入、越狱、内容风险、敏感数据、密钥及权限，并决定智能体可以连哪个大模型、用哪些工具、以什么身份访问哪些系统，同时承担 Token 计数与预算配额管理。端侧通过 SDK、Plugin 和客户端安全伴侣嵌入智能体内部，采集工具参数、文件与命令、进程及网络行为，天擎智能体管理模块识别与管控员工终端上私自安装的第三方智能体，对高风险调用实施隔离执行和外发阻断。

三是关联意图与实际行为。SDK 引出的行为数据与 AI 网关日志、模型调用告警、SWG 日志等统一汇聚分析，两类数据以会话和智能体身份关联，由流量侧判断“想做什么”，端侧确认“实际做了什么”，识别意图与行为不一致、多轮渐进诱导以及跨智能体级联风险，弥补单一网关无法观察端侧执行、单一端点缺少完整语义上下文的盲区；发现智能体“干坏事”即可定位处置、失控即“关禁闭”，并结合零信任把智能体身份绑定到人和设备，谁的智能体谁负责，出事可溯源、可追责。

四是建立“看得见、拦得住、检得出、查得清”能力闭环：持续识别模型、智能体、接口、密钥、终端、工具、插件和用户，建立影子 AI 清单、权限画像与风险等级；以确定性静态策略为底线，结合任务意图、上下文和运行环境动态调整工具权限，对高风险操作实施确认、降权、隔离或阻断，互联网下载的智能体一律禁入内网、违规访问可拦可控，智能体 skill 应用经 SafeSkill 安全商

店统一托管分发、从源头防“毒技能”；通过规则、意图识别、上下文关联和行为基线分层研判已知模式与未知变体，并将会话、调用、进程和告警关联为攻击链，支持回放、画像、历史追踪和合规报告，对接 AISOC 运营实现全局统一与自动化运营，并通过 AI 安全开放平台持续获得智能体安全赋能。

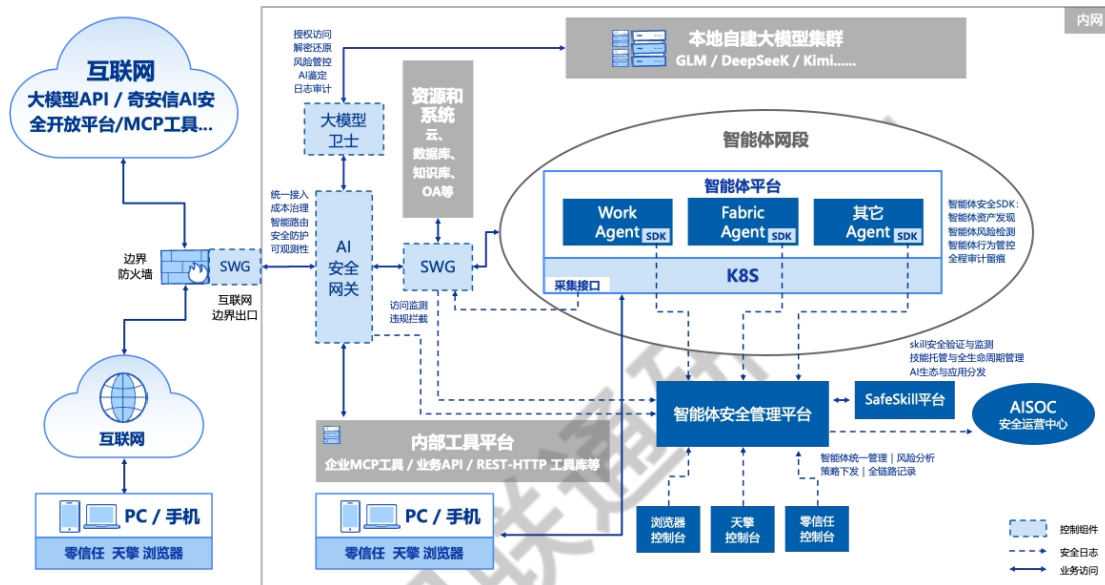


图 5-10 奇安信智能体安全解决方案总体架构

六、AI 安全治理发展建议

立足人工智能安全治理体系建设目标，本白皮书从法律法规、标准体系、产业生态、技术攻关、人才培养等多个维度提出人工智能安全治理发展建议，推动我国人工智能技术健康、安全、可持续发展。

（1）完善法律法规及治理体系

建议国家加快出台人工智能安全专项法律法规，明确 AI 研发、部署与使用等全生命周期各环节的责任主体、责任要求与监管边界，强化对数据滥用、算

法歧视、深度伪造等高风险行为的法律规制。建立分级分类的安全审查机制与动态风险预警体系，面向电信、金融、医疗、交通等重点垂直领域，构建场景化风险分析框架与合规指引。

（2）健全全流程 AI 安全标准体系

加快构建覆盖数据、算法、模型、应用及基础设施的全流程 AI 安全标准体系，重点围绕智算底座安全、模型训练与推理可信性、生成内容可追溯性、自主 AI 智能体安全等议题，制定可量化、可验证的技术规范与评估指标，并参考国内外先进实践，推动具备行业共识的通用安全基线与分领域安全应用指南，促进 AI 技术研发与产品落地的规范化与互操作性。

（3）推动产业协同与生态共建

借鉴网络安全现代产业链的建立模式，鼓励运营商、设备商、云服务商与安全企业联合构建 AI 安全共治生态，推动建立跨企业、跨平台的安全信息共享机制，完善威胁情报共享与联动响应体系，引导企业在 AI 产品设计阶段中内嵌安全能力。同时，通过生态协同强化对开源模型、第三方组件等供应链环节的安全评估与认证，提升整个产业链的抗风险能力。

（4）开展前沿安全技术攻关

聚焦 AI 安全核心技术瓶颈，系统开展模型鲁棒性增强、对抗攻击检测与防御、提示词注入防护、大模型安全对齐与幻觉抑制、AI 智能体安全防护等关键技术攻关。构建覆盖研发、测试、部署、运行全生命周期的 AI 安全监测与防护体系，推动新型防护机制与现有安全能力协同，为 AI 系统提供端到端的安全保

障。

（5）强化复合型人才培养与产学研协同

加强 AI 安全交叉学科建设，培育兼具 AI 算法开发能力与安全防御思维的复合型人才。鼓励企业与科研院所共建联合实验室、攻防演练平台和测试验证环境，加速 AI 安全评估工具、语料库及红蓝对抗技术的成果转化。通过专项培训、技能竞赛和能力认证，提升技术人员的安全意识与实战能力，建设多层次、可持续的 AI 安全人才梯队。



中国联通研究院

缩略语

缩略语	英文全称	中文释义
AI	Artificial Intelligence	人工智能
AGI	Artificial General Intelligence	通用人工智能
DL	Deep Learning	深度学习
LLM	Large Language Model	大语言模型
MLLM	Multimodal Large Language Model	多模态大语言模型
UMFM	Unified Multimodal Foundation Model	统一多模态基础模型
RAG	Retrieval-Augmented Generation	检索增强生成
API	Application Programming Interface	应用程序接口
AIGC	Artificial Intelligence Generated Content	人工智能生成内容
CoT	Chain of Thought	思维链
MCP	Model Context Protocol	模型上下文协议
A2A	Agent to Agent Protocol	代理对代理协议
APP	Application	应用程序

参考文献

- [1] Vaswani A , Shazeer N , Parmar N , et al. Attention is all you need[C]//Advances in Neural Information Processing Systems: v.30. 2017.
- [2] Rombach R , Blattmann A , Lorenz D , et al. High-resolution image synthesis with latent diffusion models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022 : 10674-10685.
- [3] Liu H , Li C , Wu Q , et al. Visual instruction tuning[C]//Advances in Neural Information Processing Systems: v.36. 2023: 34892-34916.
- [4] Zhao S , Zhang X , Guo J , et al. Unified multimodal understanding and generation models: advances, challenges, and opportunities[PP]. arXiv (2026-01-26) .
- [5] Ma X , Zhang Y , Fu Y. Towards unified multimodal large language models: a survey[C]//Findings of the Association for Computational Linguistics, 2026: 37212-37230.
- [6] Xie J , Mao W , Bai Z , et al. Show-o: one single transformer to unify multimodal understanding and generation[C]//International Conference on Learning Representations: v.2025. 2025: 28240-28264.
- [7] Xi Z , Chen W , Guo X , et al. The rise and potential of large language model based agents: a survey[J]. Science China Information Sciences ,

2025, 68 (2) : 121101.

[8] Wooldridge M, Jennings N R. Intelligent agents: theory and practice[J].

Knowledge Engineering Review, 1995, 10 (2) : 115-152.

[9] Wei J , Wang X , Schuurmans D , et al. Chain-of-thought prompting elicits reasoning in large language models[C]//Advances in Neural Information Processing Systems: v.35. 2022: 24824-24837.

[10] Schick T, Dwivedi-Yu J, Dessi R, et al. Toolformer: language models can teach themselves to use tools[C]//Advances in Neural Information Processing Systems: v.36. 2023: 68539-68551.

[11] Yao S, Zhao J, Yu D, et al. ReAct: Synergizing Reasoning and Acting in Language Models[C]//International Conference on Learning Representations, 2023.

[12] Yao S, Yu D, Zhao J, et al. Tree of Thoughts: Deliberate Problem Solving with Large Language Models[C]//Advances in Neural Information Processing Systems: v.36. 2023: 11809-11822.

[13] Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks[C]//Advances in Neural Information Processing Systems: v.33. 2020: 9459-9474.

[14] Liu Y, Chen W, Bai Y, et al. Aligning Cyber Space With Physical World: A Comprehensive Survey on Embodied AI[J]. IEEE/ASME

Transactions on Mechatronics, 2025, 30 (6) : 7253-7274.

[15] Zitkovich B, Yu T, Xu S, et al. RT-2: vision-language-action models transfer web knowledge to robotic control[C]//Proceedings of The 7th Conference on Robot Learning: v.229. PMLR, 2023: 2165-2183.

[16] Jiao R, Xie S, Yue J, et al. Can We Trust Embodied Agents? Exploring Backdoor Attacks against Embodied LLM-Based Decision-Making Systems[C]//International Conference on Learning Representations: v.2025. 2025: 86638-86668.

[17] 人工智能安全研究报告--技术视角下的安全风险梳理与应对（2025 年）[R]. 中国信息通信研究院人工智能研究所, 2025.

[18] 人工智能安全治理研究报告——推进人工智能安全治理产业实践框架（2025 年）[R]. 中国信息通信研究院, 2026.

[19] Díaz S, Kern C, Olive K. Google' s Approach for Secure AI Agents: An Introduction[R]. Google, 2025.

[20] AI 安全解决方案白皮书——构建面向 AI 基础设施的安全架构[R]. 华为云计算技术有限公司, 2026.

[21] 前沿人工智能风险管理框架[R]. 上海人工智能实验室、安远 AI, 2026.

[22] 人工智能安全风险测评（2025 年）白皮书[R]. 中国信息安全测评中心, 2026.

[23] 生成式 AI 安全白皮书[R]. 火山引擎, 2025.

- [24] 黄河燕, 李思霖, 兰天伟等. 大语言模型安全性: 分类、评估、归因、缓解、展望[J]. 智能系统学报, 2025, 20 (1) : 2-32.
- [25] 何新磊, 徐国文, 韩星烁等. 人工智能安全与隐私综述[J]. 中国科学:信息科学, 2026, 56 (2) : 481-482.
- [26] 贾堃, 张钰歆, 陈继昀等. 面向安全治理的大语言模型风险分析与应对策略研究[J]. 中国工程科学, 2026, 28 (2) : 97-112.
- [27] OWASP Top 10 for LLM Applications 2026[R]. OWASP GenAI Security Project, 2026.
- [28] Liu X , Zhu Y , Lan Y , et al. Safety of Multimodal Large Language Models on Images and Text[C]//Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, 2024: 8151-8159.
- [29] Sahoo P , Meharia P , Ghosh A , et al. A comprehensive survey of hallucination in large language, image, video and audio foundation models[C]//Findings of the Association for Computational Linguistics, 2024: 11709-11724.
- [30] OWASP Top 10 for Agentic Applications 2026[R]. OWASP GenAI Security Project, 2025.
- [31] State of Agentic AI Security and Governance[R]. OWASP GenAI Security Project, 2026.
- [32] 中国联通自智网络白皮书（2025）[R]. 中国联合网络通信有限公司,

2025.

[33] 中兴通讯自智网络白皮书（2025）[R]. 中兴通讯股份有限公司, 2025.

[34] Xu H, Wang S, Li N, et al. Large language models for cyber security: a systematic literature review[J]. ACM Transactions on Software Engineering and Methodology, 2025.

[35] Zhang J, Bu H, Wen H, et al. When LLMs meet cybersecurity: a systematic literature review[J]. Cybersecurity, 2025, 8 (1) : 55-95.

[36] Liu J, Wang K, Chen Y, et al. Large language model-based agents for software engineering: a survey[J]. ACM Transactions on Software Engineering and Methodology, 2026.

[37] Chen Y, Sun W, Fang C, et al. Security of language models for code: a systematic literature review[J]. ACM Transactions on Software Engineering and Methodology, 2026, 35 (7) : 1-65.

[38] Devlin J, Chang M-W, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019 : 4171-4186.

[39] Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge[J]. Nature, 2023, 620 (7972) : 172-180.

- [40] Kim Y, Park C, Jeong H, et al. MDAgents: an adaptive collaboration of LLMs for medical decision-making[C]//Advances in Neural Information Processing Systems: v.37. 2024: 79410-79452.
- [41] Thirunavukarasu A J, Ting D S J, Elangovan K, et al. Large language models in medicine[J]. Nature Medicine, 2023, 29 (8) : 1930-1940.
- [42] Brown T B, Mann B, Ryder N, et al. Language models are few-shot learners[J].
- [43] Wang W, Ma Z, Wang Z, et al. A survey of LLM-based agents in medicine: how far are we from baymax?[C]//Findings of the Association for Computational Linguistics, 2025: 10345-10359.
- [44] 张凌寒. 生成式人工智能的法律定位与分层治理[J]. 现代法学, 2023, 45 (4) : 126-141.
- [45] 人工智能安全治理框架 2.0[R]. 全国网络安全标准化技术委员会、国家计算机网络应急技术处理协调中心, 2025.

中国联通研究院根植于联通集团（中国联通直属二级机构），作为中国联通科技创新专业子公司，自2022年起，挂牌成立中国网络安全研究院、下一代互联网宽带业务应用国家工程研究中心及首个国家级网络安全产业知识产权运营中心，形成“两院两中心”发展格局，开创了研究院高质量发展的新局面。

中国联通研究院作为服务于国家战略、行业发展、企业生产的战略决策参谋者、技术发展的引领者、产业发展的助推者，坚持以高水平科技自立自强为使命担当，聚焦网络强国、数字中国主责，拓展联网通信、算网数智业务，形成“态度、速度、气度、有情怀、有格局、有担当”的企业文化，以下一代互联网、光网络、5G-A/6G、网络安全、数智网优、低空智能网联和新型智库研究七个领域为主攻方向，坚持“四个聚焦”，开展关键核心技术攻关、科创力量建设、专业技术人才队伍建设、创新成果转化等工作，着力实现价值创造提升、战略科技力量提升、专精特新能力成色提升，争做通信行业科技创新主力军，努力建设成为“国家信赖、行业领先、集团倚重、员工自豪”现代化一流研究院。

战略决策的参谋者 技术发展的引领者 产业发展的助推者

态度、速度、气度

有情怀、有格局、有担当

中国联合网络通信有限公司研究院

地址：北京市亦庄经济技术开发区北环东路1号

电话：010-87926100

邮编：100176



中国联通研究院



中国联通泛终端技术