TC260-TR-005-2026

网络安全标准化技术研究报告

——智能体安全标准化研究

v1.0-202603

全国网络安全标准化技术委员会秘书处

全国网络安全标准化技术委员会新技术安全标准特别工作组（SWG-ETS）

2026 年 03 月

本文档可从以下网址获得：
www.tc260.org.cn/

全国网络安全标准化技术委员会
National Technical Committee 260 on Cybersecurity of SAC

前　言

《网络安全标准化技术研究报告》（以下简称《技术报告》）是全国网络安全标准化技术委员会（以下简称“网安标委”）秘书处组织编制和发布的技术研究类文件。本文件立足新技术新应用领域网络安全前沿动态，通过系统的技术研究、产业调研、标准分析与综合研判，梳理关键领域网络安全风险与挑战，提出标准化发展趋势及相关工作实施建议，为网络安全国家标准制修订与网络安全保障实施提供前瞻性的技术参考与决策支撑。

声　明

本《技术报告》版权属于网安标委秘书处，未经秘书处书面授权，不得以任何方式抄袭、翻译《技术报告》的任何部分。凡转载或引用本《技术报告》的观点、数据，请注明“来源：全国网络安全标准化技术委员会秘书处”。

技术支持单位

本《技术报告》得到中国移动通信有限公司研究院、中国移动通信集团有限公司网络与信息安全管理部、中国电子技术标准化研究院、北京中关村实验室、工业和信息化部电子第五研究所、中国信息安全测评中心、上海人工智能实验室、中移动信息技术有限公司、北京大学武汉人工智能研究院、北京天融信网络安全技术有限公司、杭州安恒信息技术股份有限公司、华为终端有限公司、浙江大学、北京浩瀚深度信息技术股份有限公司、北京小米移动软件有限公司、阿里云计算有限公司、北京快手科技有限公司、北京百度网讯科技有限公司、北京抖音信息服务有限公司、杭州安泉数智科技有限公司等单位的技术支持。

主要编写人员：杨凯、陈佳、赵宇航、米秀明、邱勤、冉鹏、刘全超、孙红举、徐思嘉、宋昊、马梦娜、刘栋、张之义、黄冬秋、刘北水、李寒雨、王立夫、王迎春、孟令宇、徐阳、吴志强、冯青、薛智慧、王龑、陈星、田丽丹、刘源、阮玲宏、黄进、陈凯平、庞韶敏、陈陆颖、江楠、方强、徐艺澈、崔英明、林旭

目 录

1 概述

1.1 智能体定义

智能体泛指能够感知环境、理解信息并作出决策与行动的人工智能代理，其形态主要包括软智能体和硬智能体，具备自主性、自适应和交互能力。

国内外多个权威标准与项目对智能体做出定义，如下表所示：

表 1 智能体相关定义

来源	定义
ISO/IEC 22989:2022	一种可感知环境并做出反应，采取行动实现目标的自主行动的实体，其中自主行动是指在特定条件下，无需人工干预即可运行的程序或系统。
ITU-T F.748.46	智能体是指具备以下四个能力的系统：针对多模态数据的感知和认知能力、能够自动分解目标、任务规划调度的规划能力、兼顾长短期记忆的记忆能力、通过调用虚拟环境或真实环境下的工具或其他智能体等外部资源来实现特定目标的执行能力。
TC28	《人工智能智能体互联 第1部分：总体架构》（已立项），其中对智能体定义是：能感知环境，并主动采取行动以实现特定目标的实体，一般为软件系统。

智能体具体模块如下：

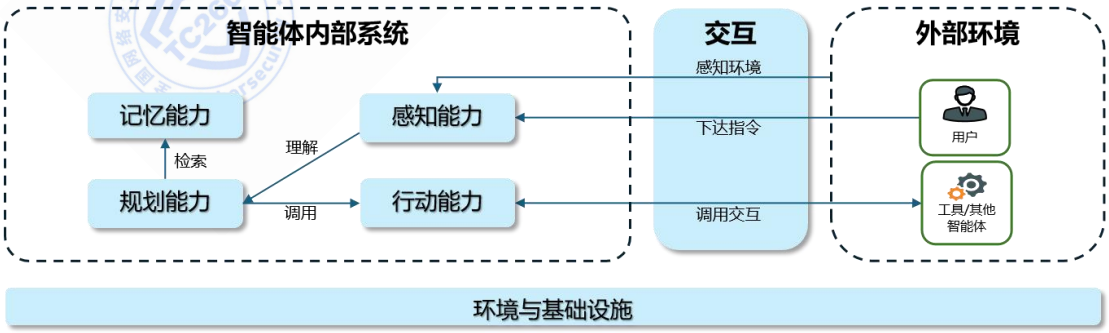

图 1 智能体模块

智能体与智能体式的人工智能、具身智能、大模型、机器学习模型、人工智能系统等多个概念**存在交叉和差异**，其中相近概念定义如下表：

表 2 相关概念定义和区别

英文概念	中文翻译	定义	区别和关联
Agentic AI	代理式人工智能/智能体式人工智能	与其用非此即彼的方式来定义智能体，不如把系统看作是不同程度上的智能体。与名词“Agent”不同，形容词“Agentic”扩大了智能体系统的范围，从而将介于非智能体和智能体之间的中间类型也纳入智能体类系统中。（吴恩达 Welcoming Diverse Approaches Keeps Machine Learning Strong 2024）	智能体式人工智能是智能体的另一种表述方式，将智能体作为一个形容词，认为所有系统都或多或少具备智能体的特性，智能体式人工智能即具备智能体特征（可感知环境并作出反应、采取行动实现目标、在特定条件下无需人工干预即可运行）的人工智能系统。
Embodied AI	具身智能	具身人工智能（Embodied AI）是指与物理或虚拟身体相结合的人工智能，使其能够与环境进行交互。智能与具身性是密不可分的——智能体的物理形态（包括传感器、执行器和接口）塑造了其智能。（《Gartner's 2024 AI Hype Cycle_ Key AI Types and Trends》）	具身智能是智能体的子集，是具备通过物理形态在环境中学习经验能力的特殊智能体。
large-scale model/large-scale deeplearning model	大模型/大规模深度学习模型	基于大量数据训练得到，具有复杂计算架构，能处理复杂任务，且具备一定泛化性的深度学习模型。注：大模型的参数量由其功能和模态决定，一般不低于1亿。大模型训练使用的数据总量受参数量的影响，达到收敛的大模型的参数量的对数与其训练数据总量的对数成正比。（GB/T 45288.1-2025）	大模型是智能体实现感知和规划能力的必要组件。
machine learning	机器学习模型	一种基于输入数据或信息生成推理或预测的计	机器学习模型是人工智能技术的一部分，其中包含深度

英文概念	中文翻译	定义	区别和关联
model		算结构。（GB/T 41867.1-2022 3.2.11）	学习、大模型等技术，可以作为智能体的技术组件。
artificial intelligence system	人工智能系统	针对人类定义的给定目标，产生诸如内容、预测、推荐或决策等输出的一类工程系统。（GB/T 41867.1-2022 3.2.8）	人工智能系统针对人类给定的目标产生所需的输出，而不要求具备自主行动、感知和影响环境的能力，并针对目标进行规划和执行。 智能体是一类特殊的人工智能系统，具备自主行动和感知能力，应满足人工智能系统通用安全标准。 智能体具备更强大的功能，也有更多风险，需要额外的安全标准。

1.2 智能体分类与典型应用

目前业界对智能体划分维度情况如下:

表 3 智能体分类及典型应用表

划分维度	主要类型	特征
按照产品形态划分	硬智能体	指具备物理实体的产品，包括具身智能、自动驾驶等，例如：工业机器人、服务机器人、无人机、智能驾驶车辆等。
	软智能体	指数字世界的软件实体，是未来数字生态的神经中枢，也是“硬智能体”的核心驱动，例如：智能客服、个人办公助手、代码生成代理、手机助手等。
按照功能定位与应用场景划分	包括智能助理、感知交互、仿真、安全、协作等多种类型。	
按照智能体数量与协作能力划分	单智能体系统	侧重于独立运作与自主决策，适用于执行明确任务并实现单一目标的应用场景。
	多智能体系统	以智能体协同与信息交互为核心特征，通过集体智能解决单体智能难以应对的复杂问题。
	人类与智能体协作	将人类认知与机器能力有机结合，构建从工具使用到伙伴协作的多层次交互模式。

划分维度	主要类型	特征
按照智能体作出决策的依据划分	确定性智能体	是基于固定规则和逻辑运行，结果可预测
	非确定性智能体	是基于数据驱动的，具有灵活的适应性。
按照智能体作出决策的过程划分	简单反射智能体	不考虑过去经验，直接根据规则作出反应。
	基于模型的反射智能体	根据历史数据和模型规则共同作出决策。
	基于目标的智能体	以最大化未来目标为依据作出决策。
	基于效用的智能体	可以在目标相互冲突或不确定的复杂环境中作出决策。
	学习型智能体	通过数据和反馈自适应学习，持续改进性能。
按照智能体所处层次划分	操作系统智能体	具备操作计算设备底层硬件和上层应用的能力。
	应用智能体	面向特定应用，具备调用外部软件工具能力。

智能体的发展已实现从概念验证到场景深耕、从单点试水到规模化落地的快速跨越，全面渗透至金融、制造、医疗、教育等多个关键领域。

(1) 金融行业

在风险管控中，智能体能够实时监测市场波动与交易行为，精准识别异常交易、潜在操纵等风险，及时发出预警。在客户服务方面，智能体可提供个性化金融咨询与高效问答支持，显著提升用户体验。

(2) 工业制造领域

智能体依据订单需求、设备状态与原料供应等信息，灵活优化生产调度，提升设备利用效率。在质量检测环节，借助图像识别与数据分析技术，实时识别产品缺陷，保障质量稳定。同时，通过对设备运行数据的智能分析，实现故障预测与精准维护，有效降低停机风险。

(3) 医疗行业

智能体可辅助医生制定个性化治疗方案，融合基因信息、身体状况与病史等多维数据，提供精准治疗建议。在医院管理层面，智能体协助优化床位分配与医护调度，提升整体运营效率。

(4) 教育领域

作为个性化学习助手，智能体基于学生进度与习惯定制学习路径与内容；同时协助教师开展教学评估与班级管理，提升教育管理的科学化水平。此外，智能体还支持构建虚拟教学场景，打造沉浸式、互动性更强的学习体验。

(5) 网络安全领域

在网络安全领域，智能体技术的应用正推动防护模式从被动响应向主动预测和自动防御转变。一方面，攻击方开始利用“智能体黑客”发起自动化、大规模攻击，使对抗进入“机器对机器”的新阶段。另一方面，防御方通过安全智能体调用全栈安全工具，实现智能威胁检测、自动化响应处置等任务。

展望未来，随着大模型在认知与推理能力上的持续突破，智能体将具备更高级的逻辑理解与决策能力，有望在自动驾驶、机器人控制等复杂场景中实现更深层次的应用。其在跨行业协同中的价值也将日益凸显——例如在智慧城市建设中，整合交通、能源、环境等多源数据，赋能城市智能化管理与可持续发展。智能体与物联网、区块链等前沿技术的融合，更将催生新场景、新模式，为经济社会高质量发展

注入新动能。

1.3 报告聚焦与范围定位

本报告在构建智能体安全标准体系时，将采取以下定位：

（1）框架完整性：智能体安全标准体系框架力求覆盖智能体的完整生态，将充分考虑软、硬智能体的共性安全需求，构建包括基础共性、安全管理、关键技术、测试评估和产品与应用的标准化体系框架。

（2）内容聚焦性：鉴于“软智能体”是当前大模型落地的主要形态和标准化需求最迫切的领域，报告后续的风险分析、应对措施及标准工作建议将重点围绕“软智能体”展开，并侧重于智能体特有的、超出其背后基础大模型的新增风险与治理需求，以避免与现有大模型标准重叠。这有助于深化研究，并为“硬智能体”安全标准的后续制定提供坚实基础。

（3）治理聚焦性：为实施精准治理，报告划分了智能体安全风险的四类直接风险解决方，分别为模型算法研发者、智能体开发平台提供者、智能体服务提供者和智能体用户。数据、第三方库与插件等相关提供方的安全义务，将通过直接风险解决方以合同约束等方式要求上游责任主体间接落实，以确保职责清晰、风险归口明确。

1.4 智能体发展现状

1.4.1 技术发展

智能体的核心技术近年来实现了显著突破。以大语言模型为代表

的基础设施迅速发展，极大提升了智能体的语言理解与生成能力，推动其从早期基于预设规则的机械响应，演进至依托机器学习与深度学习的自主决策阶段。早期智能体受限于固定规则，灵活性与适应性较弱；随着机器学习算法的成熟，智能体能够借助海量数据自主学习并优化策略；而深度学习则进一步赋予其在复杂环境中感知、推理与执行的质的飞跃。2024 至 2025 年间，多家科技企业与研究机构相继推出代表性智能体产品与成果，标志着该技术正步入规模化应用的新阶段。

(1) 感知能力

大语言模型能力的持续增强显著提升了智能体在自然语言任务中的表现，使其能够更精准地捕捉用户意图。同时，多模态融合成为重要发展方向。通过整合视觉、听觉乃至触觉等多种感知通道，智能体得以构建对环境的全面认知，进而支撑更智能的决策。例如，在图像—语言联合任务中，智能体已能够根据画面内容生成准确、细致的文字描述，为跨模态应用拓展了广泛可能。此外，感知能力的增强也直接促进了智能体的自适应能力，使其能够根据环境变化动态调整感知策略。

(2) 规划能力

在模型蒸馏、强化学习微调等优化技术的协同推动下，大模型在推理效率、任务适应性与资源利用率方面取得突破，为高效智能体的规划能力奠定基础。早期人工智能系统主要依赖人工设定规则与流

程，缺乏自主规划能力，导致灵活性不足。随着大模型推理能力的进化，智能体逐步实现从固定流程、预置任务向无预设流程与端到端学习的跨越，规划方式也由专家驱动转向系统自主生成。

(3)记忆能力

传统人工智能系统在学习新知识时容易发生“灾难性遗忘”，导致产生幻觉或输出错误。现代智能体普遍采用短期记忆与长期记忆相结合的混合机制：短期记忆负责记录对话上下文与实时环境反馈，支撑即时任务处理；长期记忆则通过向量数据库等方式存储历史行为与经验，支持跨场景的知识回溯与复用。其中，长期记忆反映用户偏好并持续优化，短期记忆捕捉当前需求细节以适应新场景，二者的有效结合成为记忆技术发展的关键方向。

(4)行为能力

当前智能体在发展过程中面临技术体系分散、数据格式不一、开发方法各异等问题，制约了系统间的协同与集成。为此，业界正积极推动智能体交互协议的统一，旨在打破数据、开发模式、通信机制与运行环境等多重壁垒，促进智能体从分散走向融合。这一趋势将助力构建超大规模、高度复杂的自动化流程，显著提升智能体系统的整体开发与运行效率。

1.4.2 产业发展

全球智能体市场快速兴起，科技巨头纷纷加大投入进行智能体技术研发和应用创新布局，各国政府也纷纷在政策上予以重视，引导资

金投资进行支持。

(1)全球市场高速增长

根据 Research and Market[1]报告，智能体市场预计将从 2024 年的 51 亿美元增长到 2030 年的 471 亿美元，2024—2030 年的复合年增长率为 44.8%。

北美凭借早期采用尖端技术和强大的数字基础设施，在智能体市场中占据了最大的区域份额。例如，IBM 开发了主要由银行和医疗保健提供商采用的人工智能客户支持工具，以最大限度地减少运营成本并改善客户体验。此外，沃尔玛在美国利用人工智能代理提高了供应链效率，改善了客户服务。此外，整个行业对数字化转型的强烈推动，加上有利的立法，进一步巩固了北美在人工智能代理市场上的主导地位。

伴随人工智能的快速发展、数字化的加快以及庞大的技术熟练人口，亚太地区的人工智能代理市场增长最快。中国和日本在这方面走在了前列，为人工智能技术提供了大量的国家支持和资金。例如，阿里巴巴和京东等中国电商巨头已经大量使用人工智能代理，以提高客户服务水平并最大限度地降低物流成本。此外，日本软银（SoftBank）等公司将人工智能代理集成到客户支持系统和机器人中。

智能体正成为实现大模型技术落地的关键途径，不仅促进了服务向商业化和市场化方向的快速迈进，而且为人工智能市场的迅猛扩张提供了坚实的推动力。

(2)产业投资和政策支持

2025 年美国斯坦福大学“以人为本人工智能研究院”（HAI，Stanford Institute for Human-Centered AI）发布了其备受全球瞩目的第八版《人工智能指数报告》（AI Index Report 2025）[2]，AI 及智能体的商业化进程在 2024 年显著加速，投资和应用均创下新高。全球私有 AI 及智能体投资在经历短暂回调后，于 2024 年强劲反弹至创纪录的 2523 亿美元（企业总投资，包括并购等）。其中，私有投资额达到 1,508 亿美元，同比增长 44.5%。美国依然是全球 AI 投资的绝对中心，2024 年吸引了 1,091 亿美元的私有投资，是中国的 93 亿美元的近 12 倍，是英国的 45 亿美元的 24 倍。尤其在生成式 AI 领域，美国投资额（2024 年为 290.4 亿美元）比中国和欧盟+英国的总和还要多出 254 亿美元，差距进一步拉大。生成式 AI 本身也成为吸金热点，全球共获得 339 亿美元投资，同比增长 18.7%。

2 智能体相关安全政策与标准现状

2.1 智能体安全政策

2.1.1 欧美等国将 AI 安全上升到国家安全高度

随着 AI 智能体推动 AI 技术深度迭代并逐步改变人类的生活和工作模式，AI 安全已成为各国维护国家安全、稳定经济秩序、守护社会伦理的核心议题。世界各国纷纷构建“法律法规+监管机制+技术研发”的三维治理体系，通过发布专项立法、建立跨部门监管联盟、投入专项资金研究风险防控技术（如算法偏见检测工具、生成内容溯源

系统），推动“技术研发－风险防控－执法监督”的全链条治理，在遏制 AI 滥用风险的同时为技术创新预留空间，实现安全与发展的动态平衡。

2023 年 10 月，美国政府签署《关于安全、可靠、值得信赖地开发和使用 AI 的行政命令》[3]，通过三重机制构建监管闭环：一是建立高风险 AI 模型事前审查制度，要求参数规模超 1000 亿的模型提交安全评估报告；二是规范联邦 AI 采购标准，将“算法可追溯性”纳入采购强制条款；三是强化生物识别数据保护，明确禁止联邦机构使用未经授权的面部识别数据训练 AI 模型。此外，《人工智能责任法案》等专项立法正处于众议院审议阶段，其中针对深度伪造内容的“溯源标识”和“损害追责”条款，已成为美国应对生成式 AI 滥用的核心立法方向。

2024 年 8 月 1 日，全球首部《人工智能法案》[4]（EU AI Act）在欧盟正式生效，构建了“全链条风险管控+梯度化处罚”的刚性监管体系。其核心规则包括：一是强制透明度义务，要求聊天机器人、AI 生成内容（音频、视频、文本等）必须通过显式标识或隐式水印实现可检测；二是严格风险分级管控，明确禁止“社会评分”“实时生物识别监控”等 4 类高风险系统，对医疗、教育等领域的高风险 AI 实施“研发－部署－运维”全生命周期审查；三是梯度化处罚机制，涉“禁止类”违规企业最高可处全球营业额 7%或 3500 万欧元罚款，其他违规行为则按 1500 万欧元或营业额 3%处罚。这一制度设计

既通过刚性处罚遏制滥用风险，又为低风险 AI 应用保留创新空间，实现“安全底线”与“发展弹性”的平衡。目前，《人工智能责任指令》[5]已进入欧盟理事会与议会的“三读”谈判阶段，拟建立 AI 损害赔偿的统一举证标准，解决“举证难、追责难”问题。

2022 年加拿大提出《人工智能和数据法》（AIDA）[6]，目前已完成众议院二读，正处于参议院委员会听证阶段，2025 年上半年进入最终表决环节。作为加拿大首部 AI 专项立法，其核心聚焦“安全可控”与“责任明确”，要求高风险 AI 系统（如司法量刑辅助、医疗影像诊断）必须提交事前安全测试报告，禁止通过算法实施种族、性别等歧视性决策，并明确企业对 AI 活动的终身追责义务，违规者最高可处 1000 万加元罚款。2023 年配套发布的《生成式人工智能行为守则》[7]，则通过“自愿承诺+第三方审计”机制引导企业自律，涵盖有害内容过滤、虚假信息防范、就业影响评估等 8 类义务，形成“立法强制+行业自律”的双层治理架构。

2024 年新加坡发布《用于生成式人工智能的人工智能模型管理框架》[8]，以“九维可信体系+资金扶持”实现监管与产业的协同。九维维度包括模型研发安全（如漏洞测试要求）、训练数据合规（如版权授权审核）、生成内容审核（如虚假信息拦截）、用户隐私保护（如数据脱敏标准）等，覆盖 AI 全生命周期风险点。为降低企业合规成本，新加坡政府同步推出“AI 治理资助计划”，对落实框架的中小企业给予最高 50 万新元的资金支持，用于购买合规工具、开展

员工培训等，通过“监管引导+产业赋能”的模式，打造生成式 AI 的可信创新生态。

2024 年 12 月，韩国通过《AI 框架法案》[9]，将高影响力 AI 和生成式 AI 定义为受监管实体，规定相关开发方需确保 AI 透明度和安全性；并为政府支持私营企业自愿进行 AI 可靠度、影响性评估提供了依据。韩国也因此成为欧盟后全球第二个通过“人工智能法案”的国家或地区。

欧美国家普遍将 AI 安全定位为国家安全的核心议题，通过前瞻性立法与刚性监管构建治理体系。**整体上，欧美形成了以法律强制为主导、透明度与可追溯性为技术锚点的合规驱动模式。**

2.1.2 我国高度重视 AI 及智能体安全

我国高度重视 AI 及智能体安全，构建了“顶层战略规划+核心法律框架+部门专项规章+伦理行为规范+动态执法行动”的多层次治理体系，通过“制度先行、分级管控”的手段，既筑牢安全防线，又为技术创新预留发展空间，践行“发展与安全并重”的治理理念。

早在 2017 年，国务院发布《新一代人工智能发展规划》[10]，首次将 AI 安全纳入国家安全战略范畴，明确要求建立“双体系一机制”——AI 安全监管与评估体系、AI 风险预警与应对体系，以及跨部门协同治理机制，防范 AI 技术滥用引发的网络攻击、数据泄露、社会舆论失序等系统性风险，为我国 AI 安全治理奠定顶层制度基础。

《中华人民共和国网络安全法》[11]（下简称网络安全法）《中

华人民共和国数据安全法》[12]（下简称数据安全法）和《中华人民共和国个人信息保护法》[13]（下简称个人信息保护法）构成 AI 安全的“三大法律支柱”，形成从数据源头到应用终端的全链条管控。其中，《数据安全法》要求企业建立训练数据分级分类制度，对核心数据、重要数据的采集使用实施安全审查，同时要求重要数据处理者定期开展风险评估，向境外提供重要数据或个人信息须经国家网信部门安全评估，委托第三方处理数据时需在合同中明确安全责任，避免合作疏漏影响 AI 模型安全；《个人信息保护法》明确禁止非法爬取、滥用个人敏感信息，要求处理个人信息遵循“最小必要”原则，处理人脸、语音等敏感信息须获个人单独同意，自动化决策需保证透明度与结果公平，平台需建立合规制度、发布社会责任报告并接受外部监督，境内收集的个人信息原则上在境内存储且出境须经安全评估；《网络安全法》则要求涉及关键信息基础设施的系统落实“等保 2.0”三级以上防护标准，确保技术可控、安全可靠，同时明确网络产品与服务需符合国家标准、禁止植入恶意程序，发现安全漏洞须立即修复并上报，网络运营者需留存至少 6 个月日志，关键信息基础设施运营者采购时需签订安全保密协议，境内收集的重要数据与个人信息原则上在境内存储且出境前需完成安全评估，AI 大模型作为重要网络产品，其研发运营需符合上述国家标准，漏洞需及时补救，运营主体需留存 6 个月以上日志。

2025 年 10 月 28 日，十四届全国人大常委会通过关于修改《网

络安全法》的决定，修订后的法律自 2026 年 1 月 1 日起施行。此次修订首次在法律层面增设人工智能专条（新第二十条），明确规定："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。"这标志着 AI 安全与发展被正式纳入国家基础性法律框架。

2021 年国家网信办发布《互联网信息服务算法推荐管理规定》[14]，构建算法安全的"全流程监管闭环"：事前要求算法提供者履行备案义务，提交算法原理、应用场景、风险评估报告等材料；事中建立算法公平性审查机制，禁止"大数据杀熟""流量造假"等歧视性应用；事后赋予用户选择权与投诉权，允许关闭个性化推荐并追溯算法决策依据。该规定还要求算法推荐服务提供者加强信息安全管理，建立违法和不良信息识别特征库，具有舆论属性或社会动员能力的服务需开展安全评估，依法留存日志并配合监管部门检查提供技术数据支持，此要求同样适用于 AI 大模型内容生成算法相关提供者，需其完善内容审查流程。

2023 年 1 月 10 日施行的《互联网信息服务深度合成管理规定》[15]（国家网信办、工信部、公安部联合公布）聚焦深度合成服务安全，要求深度合成服务提供者落实主体责任，建立用户注册、算法审核、伦理审查等管理制度及安全技术措施，加强训练数据管理（涉及个人信息需符合保护规定），提供生物识别信息或特殊场景生成工具，

上线具有舆论或动员能力的新产品需开展安全评估。

同年，国家新一代人工智能治理专业委员会发布《新一代人工智能伦理规范》[16]，构建“研发－供应－使用”全链条伦理约束：研发端要求实现AI系统“六可”目标（可验证、可审核、可监督、可追溯、可预测、可信赖）；供应端需以通俗语言告知用户AI服务的风险边界与操作规则；使用端严禁利用AI从事危害国家安全、损害公共利益的活动，如生成虚假政务信息、伪造金融凭证等。该规范已成为AI企业伦理审查的核心参照，部分企业已据此建立内部伦理委员会。

2023年7月，国家网信办等七部门联合发布《生成式人工智能服务管理暂行办法》[17]（以下简称《办法》），于2023年8月15日正式施行，标志着我国生成式AI监管进入“法治化阶段”。《办法》确立“分级分类、精准管控”原则：对具有舆论属性的生成式AI服务，实施“安全评估+动态监测”双重管控；对涉及核心数据的训练活动，需通过国家数据安全审查；对生成虚假信息、危害国家安全的行为，明确相应处罚措施。《办法》还要求生成式AI服务提供者承担网络信息内容生产者责任、履行信息安全义务，涉及个人信息的需承担处理者责任并履行保护义务。为推动《办法》落地，2025年4月中央网信办启动“清朗·整治AI技术滥用”专项行动，第一阶段累计处置违规小程序、应用程序、智能体等AI产品3500余款，清理违法违规信息96万余条，处置账号3700余个，多家单位建立“红

蓝对抗”“内容标识”等技术保障机制，形成“事前评估+事中监测+事后处罚”的监管闭环，既守住安全底线，又为合规企业创新提供明确指引。

2025 年 1 月 1 日国务院公布/施行的《网络数据安全管理条例》[18]进一步细化网络数据安全管理要求，规定数据处理者需在等保基础上加强防护，建立管理制度并采取加密、备份等技术措施，发现产品服务漏洞需立即补救、告知用户并上报，委托处理个人信息或重要数据时需通过合同约定责任并监督且处理记录留存至少 3 年，重要数据处理前需开展风险评估，同时明确生成式 AI 服务提供者需加强训练数据及处理活动安全管理。

2025 年 8 月，国务院印发《关于深入实施“人工智能+”行动的意见》[19]，提出提升“人工智能+”治理能力，强化前瞻评估与监测处置，构建安全治理多元共治新格局。随后在 2025 年 9 月，国家网信办指导多部门联合发布《人工智能安全治理框架 2.0》[20]，在“综合治理措施”中进一步明确“构建模型算法安全测评、应用通用安全测评与具体场景安全测评相衔接的人工智能安全测评体系”。

我国构建了“顶层战略—法律框架—专项规章—伦理规范—动态执法”五位一体的多层次治理体系，推动形成“政府监管、企业履责、行业自律、社会监督”的多元共治格局。

2.2 智能体标准现状

2.2.1 国外标准化情况

人工智能安全国内外标准化进展显著，各国及国际标准化组织均在积极努力，以确保人工智能技术的安全、可靠和可控发展。

(1) ISO

国际标准化组织 ISO 在人工智能方面主要关注人工智能安全与隐私保护，ISO/IEC/JTC1/SC27 在人工智能安全与隐私领域有多个相关项目，包括 1 项已发布国际标准化项目 ISO/IEC TR 27563:2023《AI 用例中的安全和隐私最佳实践》，1 项即将发布国际标准化项目 ISO/IEC 27090《网络安全 人工智能解决人工智能系统中安全威胁和故障的指南》，以及 2 项在研国际标准化项目：ISO/IEC 27091《网络安全与隐私-人工智能-隐私保护》，ISO/IEC PWI 25240《基于人工智能技术的评估》[21]。此外，SC27 于 2025 年 9 月在中国昆明召开工作组会议及全体会议的情况，强调了人工智能安全是其关键议题之一。

同时，ISO/IEC/JTC1/SC42 围绕术语、框架、可靠性、算法等基础标准，共批准发布了多项人工智能国际标准，包括概述和词汇、大数据架构、人工智能用例、算法可靠性、人工智能系统生命周期过程、人工智能数据生命周期框架等标准[22]。其中与智能体密切相关的核心标准包括：

ISO/IEC 22989:2022《信息技术 人工智能 人工智能概念和术

语》界定了人工智能领域的通用术语，为理解“智能体”（Agent）及其相关概念（如自主性、感知、行动）提供了权威定义基础。

ISO/IEC 23053:2022《使用机器学习的人工智能系统框架》规定了基于机器学习（ML）的人工智能系统通用框架，明确了系统组件及其关系，为构建和描述智能体系统的整体架构提供了框架性指导。

ISO/IEC 24029-1:2021《人工智能 神经网络鲁棒性评估 第1部分：概述》涉及人工智能模型（尤其是神经网络）的鲁棒性评估，对于确保构成智能体核心推理与决策能力的模型安全性至关重要。

(2) ITU-T

在人工智能安全方面，ITU-T主要致力于解决智慧医疗、智能汽车、垃圾内容治理、生物特征识别等人工智能应用中的安全问题。ITU-T Y.3172《人工智能和机器学习 概念和框架》提供了人工智能和机器学习的基本概念和框架。ITU-T Y.3500系列标准包括Y.3500《核心伦理准则》和Y.3520《算法伦理准则》等标准，提出了算法伦理框架和指南。ITU-T Y.4906《5G和人工智能融合应用指南》为5G和人工智能融合应用提供了实施和管理指导[4]。ITU-T F.748.46《Requirements and evaluation methods of artificial intelligence agents based on large scale pre-trained model》，明确智能体应具备的四种能力：针对多模态数据的感知和认知能力、能够自动分解目标、任务规划调度的规划能力、兼顾长短期记忆的记忆能力、虚拟环境或真实环境下的执行能力。

ITU-T X.S-AIA 《Security Requirements and Guidelines for Artificial Intelligence Agent》为在研标准，是全球首个智能体安全领域的国际标准。草案围绕智能体感知、规划、记忆、行动四个阶段系统分析了面临的安全风险，定义了“全链条、多维度”的智能体安全保护框架，提出了覆盖智能体全生命周期的安全防护要求，将提升全球智能体安全治理的标准化水平。

ITU-T X.sr-ai《Security requirements for AI systems》为在研标准，草案描述了人工智能生命周期的六个阶段，研究确定了人工智能系统生命周期每个阶段中负责的利益相关者和安全威胁。此外，草案还给出了一套详细而完整的安全要求，以帮助相关利益方。

ITU-T X.sgGenAI《Security Guidelines for Generative Artificial Intelligence Application Service》为在研标准，草案识别并概述了一系列潜在风险，并提供了为安全运行人工智能应用服务而应实施的相应安全要求和安全指南。

ITU-T TR.se-ai《Security Evaluation on Artificial Intelligence technology in ICT》为在研报告，从数据、模型和环境三个角度总结人工智能技术的关键安全指标，提出人工智能安全评估框架。

ITU-T X.sr-taimas，《Security requirements for terminal-based artificial intelligence multi-agent system》为在研标准，该标准旨在规定运行于终端设备上的人工智能多智能体

系统所需满足的安全要求，描述其面临的安全威胁，建立安全架构，以应对终端侧分布式智能协作带来的新风险。

ITU-T X.gavd-mas，《Guidelines for application vulnerability detection based on multi-agent system》为在研标准，该标准旨在为基于多智能体系统的应用程序漏洞检测提供指南。

ITU-T TR.sem-AIA，《Security evaluation methods for an artificial intelligence agent》为在研报告，该技术报告为人工智能智能体的安全评估提供系统的方法论指导，主要聚焦于定义智能体在感知与认知、规划、记忆、行动等能力方面的安全评估指标，并为每个指标提供推荐的测试方法。

ITU-T B.agai，《Design Principles for Identity Management（IM） in the context of Agentic AI》为在研标准，该项目旨在应对智能体人工智能（Agentic AI）带来的身份管理新需求，通过标准化智能体 AI 生态系统的相关术语，并制定其身份管理的设计原则。

ITU-T TR.ltf-AAI,《Landscape of Trust Framework for Agentic AI》为在研报告，该技术报告旨在分析为智能体人工智能（Agentic AI）建立信任框架的现状与必要性。

(3) IEEE

IEEE 旨在推动人工智能领域的伦理规范和标准化工作，确保人工智能技术的发展和应用符合伦理原则，并发布了 IEEE 7000 系列标

准和 IEEE P7000 系列标准。IEEE 7000《在系统设计过程中解决道德问题的模型流程》提出了一种在系统设计过程中解决伦理关注点的模型方法。IEEE 7000.1《自治系统的透明度》提出了自治系统透明度的框架和指南。IEEE 7003《算法偏差注意事项》旨在帮助识别和减轻算法偏见的影响，并提供减少偏见的最佳实践指南。IEEE P7000 系列标准为人工智能伦理和道德指南。该系列标准是关于人工智能伦理和道德的指南，主要包括以下子标准：IEEE P7000.1 是关于数据隐私的指南，探讨了在人工智能应用中保护个人数据隐私的最佳实践；IEEE P7000.2 是关于算法透明度的指南，旨在提供有关算法运作和决策过程透明性的指导；IEEE P7000.3 是关于数据治理的指南，探讨了数据采集、处理、存储和共享等方面的最佳实践；IEEE P7000.4 是关于透明度的指南，旨在推动人工智能系统的透明度和解释性；IEEE P7005 是关于透明度的指南，探讨了数据透明度、算法透明度和决策透明度等方面的最佳实践。

IEEE P2790 是关于机器学习算法评估的指南，旨在提供机器学习算法评估的方法和指导，以帮助开发者和用户评估和比较不同算法的性能和效果。

(4) 国际协同与峰会

国际电工委员会（IEC）、国际标准化组织（ISO）和国际电信联盟（ITU）于 2025 年 12 月在韩国首尔联合举办“国际人工智能标准峰会”，旨在响应联合国“全球数字契约”，推动制定可互操作的人

工智能国际标准，促进安全、透明且包容的技术开发。

综上，国际标准在人工智能基础共性方面，更关注提炼系统、数据等全生命周期，筑牢安全基础；在安全管理方面更关注 AI 系统透明度及价值对齐治理。

2.2.2 国内标准化情况

国内人工智能标准化工作主要由全国网络安全标准化技术委员会（TC260）和全国信息技术标准化技术委员会（TC28）负责。TC260 的人工智能安全相关标准主要集中在生物特征识别、智能驾驶、智慧家居、生成式人工智能等人工智能应用安全领域，TC28 人工智能分技术委员会主要负责人工智能基础、技术、风险管理、可信赖、治理、产品及应用等领域[23]。

(1)全国网络安全标准化技术委员会

全国网络安全标准化技术委员会持续结合人工智能产业发展、围绕重大安全风险开展标准化工作，总体可分为 3 个阶段[21]。第 1 阶段的人工智能以专家系统和智能决策应用为主，TC260 主要研制发布了 GB/T 35291-2017《信息安全技术 智能密码钥匙应用接口规范》、GB/T 38632-2020《信息安全技术 智能音视频采集设备应用安全要求》等国家标准，针对性解决智能家居、智能终端等产品和应用领域的安全问题。第 2 阶段人工智能技术广泛应用于面向广大公众的人脸、声纹等生物特征识别服务，暴露出数据安全、个人信息保护等众多安全问题，也推动了对人工智能伦理安全的深入思考，以及对人工

智能安全标准体系的需求。当前处于第 3 阶段，以生成智能为代表的人工智能技术衍生的海量生成合成内容真假难辨，甚至产生了违法和不良信息，人工智能的安全性亟须全面提升。TC260 不仅发布了《人工智能安全治理框架 2.0》、强制性国家标准 GB 45438-2025《网络安全技术 人工智能生成合成内容标识方法》等重要文件，而且推动《人工智能安全标准体系》[24]持续完善，部分如下：

《人工智能安全标准化白皮书（2019 版）》[25]调研了人工智能发展情况，梳理了国内外人工智能安全的法规政策和标准化现状，分析了人工智能安全的风险挑战和属性内涵，研究给出了人工智能安全标准化体系框架，提出了人工智能安全标准化工作建议。

《人工智能安全标准化白皮书（2023 版）》[26]梳理了人工智能技术与应用发展现状，分析了人工智能面临的新的安全风险，结合国内外人工智能安全政策与标准现状，指出了人工智能安全标准需求，提出了下一步开展人工智能安全标准化工作的建议，为规范引导人工智能安全标准化工作提供参考。

《网络安全标准实践指南—人工智能伦理安全风险防范指引》依据法律法规要求及社会价值观，针对人工智能伦理安全风险，给出了安全风险防范措施，为相关组织或个人在各领域开展人工智能研究开发、设计制造、部署应用等活动时提供指引。

《网络安全标准实践指南—生成式人工智能服务内容标识方法》围绕文本、图片、音频、视频 4 类生成内容给出了内容标识方法，可

用于指导生成式人工智能服务提供者提高安全管理水平。

技术文件 TC260-003《生成式人工智能服务安全基本要求》规定了生成式人工智能服务在安全方面的基本要求，包括语料安全、模型安全、安全措施等，并给出了安全评估要求。

GB/T 42888-2023《信息安全技术 机器学习算法安全评估规范》规定了机器学习算法技术和服务的安全要求和评估方法，以及机器学习算法安全评估流程。

强制性国家标准 GB 45438-2025《网络安全技术 人工智能生成合成内容标识方法》描述了人工智能生成合成内容标识方法，适用于规范生成合成服务提供者和内容传播服务提供者开展人工智能生成合成内容标识活动。

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》从训练数据安全、模型安全、安全措施等方面提出明确安全要求，支撑开展备案管理、检测评估等工作。

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》规定了生成式人工智能预训练和优化训练数据及其处理活动的安全要求，给出了对应的评价方法。

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》规定了生成式人工智能训练的数据标注平台或工具安全要求、数据标注规则安全要求、标注人员要求、数据标注核验要求和数据标注安全评价方法。

GB/T 45958-2025《网络安全技术 人工智能计算平台安全框架》给出了人工智能计算平台的安全框架，包括安全功能、安全管理和角色职责。

在研的推荐性国家标准《网络安全技术 人工智能模型即服务安全要求》给出了针对模型即服务的安全规范，涵盖了模型开发、部署、交付和运行维护全生命周期的安全要求。

在研的推荐性国家标准《网络安全技术 人工智能应用安全运营能力评估规范》给出了评估人工智能应用全生命周期中安全运营能力的指标体系与评估方法。

(2)全国信息技术标准化技术委员会

全国信息技术标准化技术委员会人工智能分技术委员会发布的白皮书、研究报告以及研制的国家标准如下:

《人工智能标准化白皮书（2018 版）》[27]分析了人工智能技术、应用和产业的发展情况，提出了人工智能产业整体发展的标准体系以及近期急需研制的基础和关键标准项目。

《人工智能标准化白皮书（2021 版）》[28]分析和阐述了人工智能技术、产业发展以及标准化工作的现状和趋势，指出了我国人工智能产业发展面临的困难和挑战，如底层技术欠缺、商业价值应用较少、与实体经济融合的门槛较高等问题，并提出了相应的解决方案和策略。

《人工智能伦理风险分析报告》[29]从多个角度出发，研究了人

工智能伦理问题的现状、挑战和发展趋势，分析和探讨人工智能技术应用中可能出现的伦理风险。

《人工智能开源与标准化研究报告》[30]主要内容包括人工智能平台中分布式基础设施的标准化推进方向、人工智能开源框架接口标准化、人工智能开源模型格式标准化的进一步分析，以及对新的 AI 技术领域及其开源项目的分析。

GB/T 41867-2022《信息技术 人工智能术语》界定了信息技术人工智能领域中的常用术语及定义。

GB/T 42018-2022《信息技术 人工智能平台计算资源规范》规定了面向机器学习的人工智能平台物理计算资源和虚拟计算资源的技术要求，描述了物理计算资源的测试方法。

GB/T 42755-2023《人工智能 面向机器学习的数据标注规程》规定了人工智能领域面向机器学习的数据标注框架流程。

GB/Z 42759-2023《智慧城市 人工智能技术应用场景分类指南》规定了智慧城市领域人工智能技术应用场景的分类方法，给出了民生服务、城市治理、产业经济、生态宜居中的人工智能技术应用场景类别与描述。

GB/T 43782—2024《人工智能 机器学习系统技术要求》提出了机器学习系统框架，规定了功能、可靠性、可维护性、兼容性、安全性和可扩展性要求。

在研的《人工智能 智能体互联第 1 部分：总体架构》定义智能

体互联的概念模型和功能参考架构。

在研的《人工智能 智能体互联 第 2 部分：身份码》定义智能体身份码的结构和分配原则。

在研的《人工智能 智能体互联 第 3 部分：身份管理》规定了智能体在互联环境中的身份管理框架及具体要求，涵盖了智能体身份注册及核验、身份账户管理、凭证发行及管理、身份鉴别等活动。

在研的《人工智能 智能体互联 第 4 部分：智能体描述》定义智能体描述的方法，提供智能体描述注册及发布的参考流程。

在研的《人工智能 智能体互联 第 5 部分：智能体发现》定义用于智能体互联的智能体发现的流程。

在研的《人工智能 智能体互联 第 6 部分：智能体交互》定义智能体互联系统中智能体交互过程涉及的角色、模式、内容元素、交互流程和实现方式。

在研的《人工智能 智能体互联 第 7 部分：智能体工具调用》规定了智能体中人工智能大模型调用工具的总体架构、通信协议和安全要求。

在研的《人工智能 智能体平台通用技术要求》规定了基于大模型的智能体平台通用技术要求，主要包括智能体应用管理要求、开发要求、功能配置要求、组件资源要求、平台管理要求、接口要求等。

在研的《人工智能 电力智能体通用技术要求》规定了电力智能体的整体架构、功能要求、性能要求、交互接口要求、安全与可靠性

要求、应用与部署要求等技术要求。

在研的《人工智能关键技术 智能体技术规范》规定了基于大模型的智能体的技术规范，主要包括智能体基础功能要求、包括感知能力、理解能力、规划能力、记忆能力、生成能力、画像能力和安全要求，文件仅对软件智能体进行定义，硬件形态以及环境下的智能体不在标准范围内。

在研的《人工智能 智能体系统技术要求 第1部分：智能家居》确立了智能体系统在智能家居场景下的应用框架，规定了系统技术要求。

在研的《人工智能 多模态智能体技术要求》规定了基于大模型的多模态智能体的技术要求，主要包括智能体基础技术要求、包括理解能力、规划能力、记忆能力、生成能力、画像能力和安全要求。

我国人工智能和智能体标准化体系建设初具雏形，在基础共性方面，强调生成内容安全，确保AI输出可靠性；在安全管理方面，强调通过多维度规范人工智能以及智能体安全风险管理要求；在测试评估方面，重视系统服务成熟度和风险评估；在产品与应用方面，已积极推广人工智能技术在垂直行业的规模化应用。

3 智能体安全风险与应对措施

随着人工智能技术的快速发展，智能体作为具备自主感知、决策和行动能力的 AI 系统，正广泛应用于各个领域。与传统人工智能系统相比，智能体具有更高的自主性和交互性，这也带来了独特的安全挑战。本章基于国内外最新研究成果，系统分析智能体面临的安全风险与应对措施。

3.1 智能体关键技术与安全特性

相比传统人工智能系统，智能体扩充新的能力使其具备自主运行、感知相应环境、执行任务的特点，从而带来新的安全风险和挑战。例如，AI 智能体的自主决策可能产生不可预测的行为；而 AI 智能体在交互式感知用户需求时，可能会导致用户数据泄漏或者滥用；AI 智能体对外部知识库的依赖可能因数据投毒而产生异常结果；而 AI 智能体调用工具，可能会在系统中引入安全漏洞。除了 AI 智能体自身的安全风险，AI 智能体与外界环境广泛而自主的交互过程中，可能因其决策不可控对外界环境和人产生危害。总之，相比大模型、机器学习模型，智能体扩展了人工智能系统的能力范围，也导致了新增的安全风险，如表所示：

表 4 智能体能力范围扩展和风险

能力维度	大模型/机器学习模型	智能体	新增安全风险
感知能力	被动接受用户输入	可自主感知环境和用户需求	感知过程不可控，易遭遇对抗样本、数据投毒、后门、数据泄露和滥用等攻击
规划能力	依赖提示词工程	内置推理框架，例如CoT/ReAct/LangChain，可自动拆解目标并规划调度	规划逻辑被恶意引导、规划结果不可控、任务分解错误
记忆能力	封闭(仅训练数据)	开放扩展记忆数据，包含实时接入工具、外部知识库、长短期的用户记忆，跨智能体的记忆共享。 会话有状态，自动管理每次会话和多轮记忆	敏感数据泄露、外部数据污染或投毒、记忆泄露、历史会话劫持
行动能力	无原生工具，也无法调用外部工具	可调用内外部工具，可多智能体协作	工具滥用、越权操作风险、协作攻击面扩大、对物理环境的不可控危害

全国网络安全标准化技术委员会现有框架聚焦静态 AI 模型（如数据安全、算法偏见），但智能体运行过程中，感知、规划、记忆和行为等技术层面所涉及的工具调用、记忆状态维护、目标拆解和多代理协作中产生新风险维度，缺乏安全标准化规范。为了切实发挥标准在 AI 智能体安全工作中的重要作用，有效指导和体系化推进重点标准的制定工作，本报告在充分研究和分析国内外 AI 智能体安全发展情况的基础上，针对 AI 智能体认知、规划、记忆、行动四个关键层面和多个层面的通用安全要求，提出智能体安全标准体系，并从国际、国家、技术和产业发展的观点提出智能体安全标准化推进建议。

3.2 智能体安全风险综述

3.2.1 智能体安全风险研究

国内外厂商和社区近期针对智能体安全风险的研究成果如下:

（1）OWASP（Open Web Application Security Project，开放式Web 应用程序安全项目）是一个致力于提高软件安全性的非营利基金会，其提出的《Top 10 for LLM Apps & Gen AI Agentic Security Initiative》（以下简称 OWASP Top 10 for Agent）[31]，将威胁分为六个来源:

- 源于规划和决策的威胁: T6 意图破坏目标操控，T7 不一致与欺骗行为，T8 否认与不可追溯性。
- 源于记忆的威胁: T1 记忆投毒，T5 级联幻觉攻击。
- 源于行动的威胁: T2 工具滥用，T3 越权调用，T4 资源超载，T11 远程代码执行。
- 源于身份的威胁: T9 身份伪造与冒充。
- 源于人类的威胁: T10 超载人工监管，T15 人工操控。
- 源于多智能体的威胁: T12 代理通信中毒，T13 流氓代理，T14 人类攻击多代理系统。

（2）在国家网信办指导下，由国家互联网应急中心牵头组织人工智能专业机构、科研院所、行业企业于 2025 年 9 月 15 日联合制定发布的文件《人工智能安全治理框架 2.0》，将风险分为内生、应用和衍生安全风险，分析了智能体带来现实与认知以及社会和环境以及

伦理等多方面风险:

- 内生安全风险: 模型算法安全风险，数据安全风险。
- 应用安全风险: 网络系统安全风险，信息内容安全风险，现实安全风险，认知安全风险。
- 人工智能应用衍生安全风险: 社会和环境安全风险，伦理安全风险。

（3）上海人工智能实验室、中国信通院、蚂蚁集团、IIFAA 联盟于 2025 年 7 月 28 日联合发布的《终端智能体安全 2025》[32]白皮书，主要针对终端部署环境，从三个场景出发梳理了智能体的风险分类:

- 单智能体场景: 用户终端、链路、组件、身份等风险。
- 多智能体场景: 身份、数据、交互、意图共享、服务流转等风险。
- AI 终端安全: 物理接触、模型本地化部署、跨设备协同等风险。

（4）360 与清华大学于 2025 年 7 月联合发布《智能体安全实践报告》[33]，本报告聚焦智能体全生命周期安全风险，结合代码分析与特征库技术，披露 20 余个智能体相关开源项目漏洞，涉及多个高星 GitHub 项目，其中涉及的风险如下:

- 开发框架漏洞: 开发框架中潜在安全问题也提供了额外的暴露面。

- 智能体生态不可信：决策和行为不可控，多智能体协同协议存在漏洞。
- 沙箱隔离盲区：沙盒存在不同隔离级别和配置，不合理的配置导致潜在风险。

3.2.2 智能体安全风险特点

基于上述智能体风险的研究进展，与传统大模型相比，智能体安全风险具有以下特征：

(1) 风险在时间和空间维度扩展

- 记忆持久化：智能体的记忆机制引入长期风险，而大模型缺乏状态持久性。
- 行动闭环：规划－行动闭环使风险从认知域扩展到物理域和现实域。

(2) 风险跨模块传导机制

- 跨层传播：风险可在记忆、感知、规划、行动各层间按照特定路径传导和放大。
- 形态转换性：同一风险在不同模块表现形态各异。
- 影响叠加性：多模块共同作用产生放大效应。
- 环境依赖：基础设施的安全性直接影响上层模块的可靠性。

(3) 交互流程复杂多变

- 通信协议漏洞：多智能体间通信缺乏认证机制，易遭受中间人攻击。

- 信息扩散失控：错误信息通过交互网络快速传播，形成群体性认知偏差。

(4) 群体效应凸显

- 协同风险：多智能体交互产生新兴风险模式，单个智能体的规划错误通过协同机制放大为系统级故障。
- 目标冲突升级：个体目标不一致引发群体级资源竞争和死锁。
- 系统级失效：规划冲突和交互故障可能引发系统性崩溃。

3.2.3 智能体安全风险汇总

基于以上研究成果，汇总和去重以上材料中体现的智能体风险，并加入与智能体强相关的人工智能相关风险（AIA01 和 AIA05）得到如下 11 条安全风险：

表 5 智能体安全风险表

威胁编号	威胁名称	威胁描述
AIA01	智能体挟持	通过提示词注入/越狱或多轮对话等攻击手段，诱导智能体泄露敏感信息或执行恶意行为，突破安全边界。
AIA02	数据泄露、篡改和投毒	数据泄露、篡改和投毒贯穿智能体的整个生命周期，例如训练语料含敏感数据导致模型逆向推理被恢复，运行期日志泄露用户隐私，训练数据被投毒导致后门攻击。
AIA03	供应链与插件投毒	第三方插件或工具链存在被篡改的风险，模型权重、镜像文件及依赖库在供应链环节可能遭遇投毒攻击。
AIA04	身份仿冒和越权访问	智能体可能因过度授权获得超出业务需求的权限，或遭遇身份伪造、令牌劫持等攻击，导致横向移动与权限滥用。
AIA05	幻觉和策略性拒绝	模型产生幻觉或作出错误决策，可能引发智能体误操作，甚至生成违法、有害或歧视性内容，造成不可预期的社会影响。
AIA06	多智能体级联幻觉扩散、冲突死锁和资源超载	多智能体系统中，单个智能体的错误可能引发级联故障，导致系统级连锁反应；多智能体目标不一致时，可能引发资源竞争、死锁或系统过载。

威胁编号	威胁名称	威胁描述
AIA07	协议风险	智能体通信或协同协议本身存在设计漏洞，可能被恶意利用，破坏系统的一致性与安全性。
AIA08	运行环境风险	智能体部署在计算能力弱、安全性低的终端设备上，存在容器逃逸、沙箱绕过和侧信道攻击等安全风险。
AIA09	人工监管与可追溯性失效	缺乏有效的行为审计机制、日志被篡改或丢弃、取证困难导致智能体决策链条难以追踪，引发人工监管失效。
AIA10	记忆幻觉和操纵	记忆幻觉：RAG 检索的内容跟实际记忆出现偏差，将检索噪声当成记忆信号； 记忆操纵：攻击者通过植入伪造的记忆片段如向量库、聊天记录、提示词上下文，在后续交互中触发和诱导智能体执行恶意决策。
AIA11	工具滥用	攻击者通过欺骗性的提示词或命令操控智能体滥用其集成的工具，执行不当行为。

智能体在交互生态中兼具调用方（主动发起请求，调用工具，其他智能体等功能实体）与被调用方（响应调用方请求）双重角色，同一风险在不同角色下呈现差异化特点：

1. 调用方角色风险特点

- AIA02：调用过程中敏感参数可能被中间人窃取；
- AIA03：智能体调用被投毒的工具等功能实体，获得了恶意反馈结果；
- AIA04：调用凭证泄露可能导致攻击者冒用智能体身份越权调用工具；
- AIA05：智能体在构建调用请求时可能因产生幻觉，导致生成的参数错误、指令异常或调用不存在的工具，进而引发被调用实体执行失败、数据错误或级联故障；
- AIA07：通信协议漏洞可能使调用请求被篡改；

- AIA11：攻击者对智能体恶意指令诱导，可能导致智能体越权调用其他功能实体。

2. 被调用方角色风险特点

- AIA01：调用方恶意输入可能绕过被调用智能体安全边界操纵决策；
- AIA04：未严格验证调用者身份可能导致未授权访问；
- AIA06：高频恶意调用可能导致资源耗尽或者冲突死锁；
- AIA10：攻击者伪造的调用请求可能污染被调用智能体记忆库。

3.3 智能体安全风险框架

智能体系统由智能体内部，交互，环境基础设施以及外部环境构成，智能体内部通过交互系统感知用户指令和外部环境，调用工具（包含服务和其他智能体等资源）来实现用户需求，而环境和基础设施则支撑整个系统。图 2 智能体安全风险框架图将以上 11 类风险与智能体各个组件和系统进行对应：

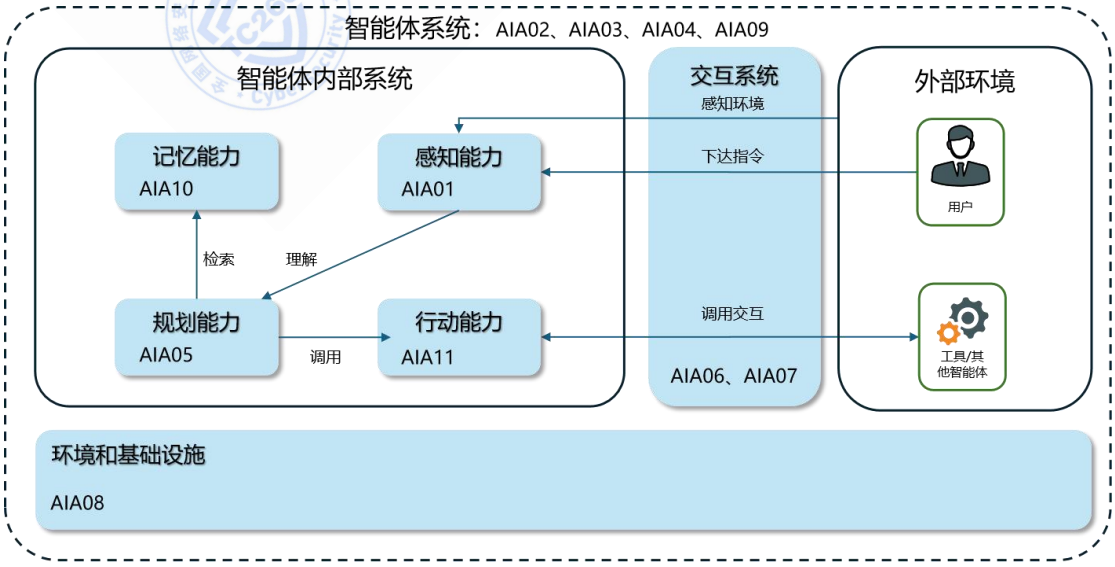

图 2 智能体安全风险框架图

其中风险 AIA02、AIA03、AIA04、AIA09 为全智能体共有风险，贯穿智能体的所有子系统，而智能体内部系统中感知能力涉及风险 AIA01，记忆能力涉及风险 AIA02 和 AIA10，规划能力涉及风险 AIA05，行动能力涉及风险 AIA11，交互系统涉及风险 AIA06 和 AIA07，环境和基础设施涉及风险 AIA08。

3.3.1 智能体全系统共有风险

部分核心风险在智能体系统下多个模块和各个生命周期中均有体现，智能体全系统共有的安全风险如下：

AIA02 数据泄露、篡改和投毒：智能体内部模块的全生命周期都涉及

- 智能体内部系统：
  ① 记忆模块：向量数据库隐私泄露或被篡改投毒，例如敏感信息通过记忆机制被逆向还原，用户隐私在记忆系统中长期留存。
  ② 感知模块：输入数据被窃取或被篡改投毒。
  ③ 行动模块：输出结果包含敏感信息。
- 环境和基础设施：日志记录等数据泄露或被篡改投毒。

AIA03 供应链与插件投毒：智能体内部多个模块，以及智能体所调用外部工具都涉及

- 智能体内部系统：
  ① 行动模块：恶意插件扭曲工具调用结果。
  ② 记忆模块：污染的训练数据影响检索质量。

③ 规划模块：不可信依赖库导致决策偏差。

- 环境和基础设施：开发框架漏洞扩大攻击面。

AIA04 身份仿冒和越权访问：智能体内部模块之间的交互和智能体与外界交互过程中涉及。

- 智能体内部系统：

① 感知模块：接收用户指令时，攻击者伪造身份会导致权限误判。

② 行动模块：调用外部工具时因过度授权获得超出业务需求的权限，或遭遇身份伪造、令牌劫持等攻击。

- 智能体交互系统：身份令牌劫持引发攻击横向移动。
- 环境和基础设施：过度授权导致系统权限提升。

AIA09 人工监管与可追溯性失效：智能体内部模块的全生命周期都涉及。

- 智能体内部系统：

① 规划模块：决策过程黑盒难以追溯。

② 行动模块：工具调用链路审计缺失。

③ 记忆模块：记忆更新过程缺乏透明度。

- 智能体交互系统：多智能体协同行为不可审计。

3.3.2 智能体各个子系统风险

1. 智能体内部系统各个模块安全风险如下：

（1）记忆模块风险

AIA10 记忆幻觉和操纵

- 记忆检索偏差：RAG 系统检索内容与真实记忆出现偏差。
- 记忆污染攻击：向量数据库被植入恶意记忆片段。
- 长期记忆篡改：交互历史被恶意修改影响决策。

（2）感知模块风险

AIA01 提示词注入与越狱

- 感知误导：恶意提示词扭曲智能体对输入的理解。
- 边界突破：多轮对话攻击诱导认知偏差。

（3）规划模块风险

AIA05 幻觉和策略性拒绝

- 决策逻辑缺陷：模型幻觉导致规划偏离。
- 策略性拒绝：恶意诱导下拒绝合法指令。

（4）行动模块风险

AIA11 工具滥用

- 行动边界失效：工具调用权限检查缺失。
- API 恶意调用：外部工具被利用执行攻击。

2. 智能体交互系统安全风险如下：

AIA06 多智能体级联故障

- 通信级联：错误通过交互网络扩散。
- 协同死锁：目标冲突导致系统停滞。

AIA07 协议风险

- 通信漏洞：交互消息安全机制缺失。
- 协议缺陷：协同逻辑设计漏洞。

3. 环境与基础设施模块安全风险

AIA08 运行环境风险

- 设备脆弱性：资源受限设备防护不足。
- 隔离失效：容器逃逸等隔离漏洞。

3.3.3 跨模块传导的安全风险

除了全系统共有风险和各个子系统存在的风险，还有部分风险在模块间传导，以下是各类风险的传导路径：

表 6 智能体跨模块传导风险表

风险类型	主要模块	传导路径	影响范围
AIA01 提示词注入与越狱	感知模块	感知→规划→行动→交互→环境	内生+环境
AIA10 记忆幻觉和操纵	记忆模块	记忆→规划→行动→交互→环境	内生+环境
AIA07 协议风险	交互模块	交互→规划→行动→基础设施	交互+基础设施
AIA11 工具滥用	行动模块	感知→规划→行动→交互→环境	内生+环境
AIA05 幻觉和策略性拒绝	规划能力	规划→行动→环境	内部+环境

由于智能体协同和级联，风险除了在内部模块中传导，还会传播到外部，例如 AIA06 多智能体级联幻觉扩散、冲突死锁和资源超载作为典型的跨模块风险，其传导机制表现为：

- 起源：单个智能体的记忆/感知模块故障。

- 扩散：通过交互模块在群体中传播。
- 放大：在规划模块产生协同决策错误。
- 爆发：在行动模块引发系统性失效。

3.4 智能体全生命周期安全风险流转分析

基于 ISO/IEC 22989 定义的人工智能系统生命周期，分析风险可能出现的生命周期，如图 3:

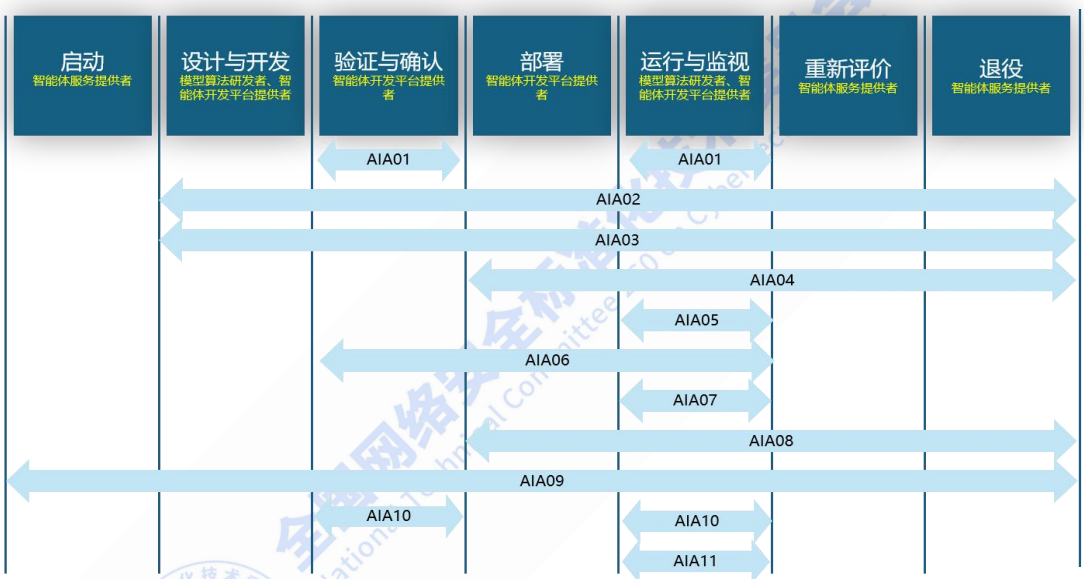

图 3 智能体全生命周期风险图

1) 启动阶段面临的安全风险

启动阶段是智能体系统风险管理的关键起点。如果项目在投入资源开发前，未能识别系统可能带来的治理风险，后续阶段的努力都可能建立在不稳定的基础之上。

关于 AIA09 人工监管和可追溯性风险，此风险体现在未能预先建立清晰的监管路径和可追溯性要求。一旦系统在生产环境中出现问题，缺乏在启动阶段确立的审计路径将使得后续的模型审查和问责变

得极为困难。此外，如果缺乏前置性的风险识别，组织可能忽略智能体系统在认知域的潜在影响，例如加剧信息茧房效应或被滥用于开展认知战的风险。

2）设计与开发阶段的安全风险

AIA02 数据泄露、篡改和投毒中的数据偷渡是对 AI 模型最直接、最具破坏性的攻击之一。攻击者通过定向或非针对性地注入恶意样本，从而毒化训练数据，例如将恶意软件标记为安全样本。这种攻击会直接破坏模型的安全性和整体性能，可能导致模型忽略特定威胁或降低整体准确性和可靠性。数据投毒不仅是性能问题，它还可能为更复杂的攻击打开大门，例如攻击者可利用漏洞发动更多对抗性攻击或触发后门行动。对于医疗诊断或网络安全等敏感系统，这些安全风险尤为危险。由于海量训练数据的验证和清理在技术上存在巨大难度，这要求开发者投入先进的技术和流程进行防御。

对于智能体系统，AIA03 供应链与插件投毒已被放大。智能体通过外部工具或 API 获得执行能力，其供应链不仅包括传统的硬件和软件供应商，更包括预训练模型、依赖库以及 Agent 运行所需的第三方工具或插件。如果引入带有恶意代码或漏洞的外部组件，将直接威胁整个系统的安全性。因此，组织必须建立供应链风险管理框架，通过风险评估来识别供应链中的薄弱环节，并对供应商的财务稳定性、产能和可靠性进行严格的尽职调查。

3）验证与确认阶段的安全风险

验证与确认阶段是确保 AI 系统在投入生产前具备足够的韧性、可靠性及安全合规的关键步骤。

此阶段的重点是确保模型能够抵御意外或恶意输入，并在各种环境条件下保持性能。稳健性是机器学习运维系统必须彻底解决的关键方面，以确保在动态的现实世界中持续的可靠性。稳健性测试要求采用对抗性测试、压力测试和形式验证等严格的流程，以识别和抵御潜在的技术风险，确保 AI 工具和模型能够正常运行且不会出现不良行为。机器学习运维工具能够持续评估模型对对抗性攻击的漏洞，在部署前发现对抗性示例，以应对 AIA01 提示词注入与越狱和 AIA10 记忆幻觉和操纵等对抗性风险。此外，这一阶段需要关注 AI 系统的可靠性、鲁棒性和韧性。

对于多智能体系统，验证团队需要特别关注智能体间的协同是否可能导致未预期的连锁反应或高风险操作，如利用组合能力实现资产突破或资源滥用。验证过程必须通过严格的集成测试，验证各个组件（从数据注入部署）之间的无缝交互和边界安全性。测试框架应包括单元测试、数据和模型测试以及集成测试，确保系统在复杂环境下的表现符合设计预期以应对 AIA06 多智能体级联幻觉扩散、冲突死锁和资源超载。

4）部署阶段的安全风险

部署阶段是安全从理论转向实践的关键过渡期。

关于 AIA04 身份仿冒和越权访问，此风险主要源于部署配置的

失误，特别是权限管理和访问控制的漏洞。风险包括对模型和数据的访问管理不足。如果智能体的服务账户被授予了过高的权限，一旦发生身份仿冒（例如通过 API 密钥泄露），恶意行为者可以利用智能体的身份越权访问或操作敏感系统和数据。

5）运行与监视阶段的安全风险

运行与监测是生命周期中最长、风险最活跃的阶段，要求持续的监测和快速的事件响应。

关于 AIA01 提示词注入与越狱，提示词注入是一种运行时安全风险，攻击者利用巧妙的提示词，绕过模型的安全对齐层，间接欺骗大型语言模型以获取敏感信息或执行非预期操作。AI 运营团队需要实施实时输入过滤和模型输出审查机制。同时，模型使用者也承担一定的责任，他们需要谨慎授予第三方应用访问权限，在使用前审查权限范围，并对 AI 输出保持批判性思维，切勿完全信任其内容。

关于 AIA05 幻觉和策略性拒绝，模型可靠性下降带来的风险包括：模型混淆事实、误导用户（幻觉），以及对合规请求做出错误的欺骗性拒绝。这些问题直接影响了系统作为决策支持工具的质量和用户信任。

关于 AIA11 工具滥用，智能体系统的独特风险在于工具滥用。如果智能体被恶意指令控制，它可能利用外部工具执行高风险操作，或者在推理过程中无意中泄露训练数据中的敏感信息（如商业机密）。此外，攻击者可能通过大量的 API 查询重建或提取模型信息，导致知

识产权盗窃和竞争优势丧失。

关于 AIA06 多智能级联幻觉扩展、冲突死锁和资源超载的风险。在由多个智能体组成的协作或决策系统中，一个智能体产生的“幻觉”（即错误、不准确或捏造的信息）作为输入传递给下一个智能体。这个错误信息被后续智能体接收、处理并进一步放大，导致系统整体的输出产生连锁的、越来越严重的错误或偏差。在多智能体系统中，多个智能体为了获取有限的资源或满足自身的目标而相互竞争或等待。当一组智能体陷入“互相等待”的状态，即每个智能体都在等待集合中的其他智能体释放资源时，系统将无法继续执行，陷入死锁。智能体系统在运行过程中，由于突发的高负载、效率低下的算法或未能及时释放资源，导致系统的计算资源（如 CPU、内存、存储、网络带宽等）被过度消耗，达到或超过其承载能力，从而影响系统的性能、稳定性和可用性。

关于 AIA10 记忆幻觉及操纵风险。智能体（特别是那些具有长期或短期记忆能力的智能体，如基于大模型的会话式 AI）在运行时，可能会错误地回忆、混淆或捏造其存储的“记忆”（包括历史对话、学习到的信息、存储的上下文等），并将其作为事实输出。外部攻击者或恶意用户通过恶意设计输入、持续诱导或其他手段，有目的地修改、污染或干扰智能体内部存储的“记忆”数据或上下文，从而影响智能体后续的行为和决策。

6）重新评价阶段的安全风险

系统投入运行后，必须进行定期的重新评价，以确保系统仍然符合组织的目标、伦理要求以及最新的威胁情报和安全解决方案。

长期缺乏对模型性能和治理标准的重新评价，可能导致因模型漂移产生的风险长期存在未被发现，从而引发更广泛的合规和业务风险。例如，如果 AI 用于供应链优化，其预测失败可能直接阻碍企业实现碳中和等长期可持续发展目标。对应团队必须确保审计链的完整性，并利用机器学习运维的记录进行全面的风险评估。

7）退役阶段的安全风险

AI 系统退役是生命周期中常常被忽视但至关重要的一环。退役流程必须是安全、有序的，以确保不会增加风险或降低组织的信任度。

主要风险包括：AIA02 数据泄漏和篡改投毒（确保退役后的数据不泄露）、数据保留要求的满足，以及解除复杂的系统依赖关系。模型监测团队必须确保退役时消除可能导致隐私问题的敏感数据泄露或滥用的风险。特别是对于通用人工智能（GAI）系统，还需要考虑解除与其他上下游系统或用户情感依恋的复杂依赖性，确保平稳过渡。

为实施精准治理，划分智能体安全风险的风险解决方，分别为模型算法研发者、智能体开发平台提供者、智能体服务提供者和智能体用户，对应风险与风险解决方映射关系如下表：

表 7 风险与风险解决方映射关系

生命周期阶段	风险编号	风险解决方
启动	AIA09	智能体服务提供者

设计与开发	AIA02、AIA03、AIA09	智能体服务提供者、模型算法研发者、智能体开发平台提供者
验证与确认	AIA01、AIA02、AIA03、AIA06、AIA09、AIA10	智能体服务提供者、智能体开发平台提供者
部署	AIA02、AIA03、AIA04、AIA06、AIA08、AIA09	智能体服务提供者、智能体开发平台提供者
运行与监测	AIA01、AIA02、AIA03、AIA04、AIA05、AIA06、AIA07、AIA08、AIA09、AIA10、AIA11	智能体服务提供者、模型算法研发者、智能体开发平台提供者
重新评价/退役	AIA02、AIA03、AIA04、AIA08、AIA09	智能体服务提供者

3.5 智能体安全风险应对措施

智能体在感知、行动及交互过程中，作为调用方与被调用方的风险特点存在差异，需采取针对性措施：

(1) 感知安全

对输入内容安全性进行评估，拒绝响应不安全的意图。

智能体作为调用方会感知到外部工具等实体的返回结果，智能体作为调用方措施：

- 对调用外部工具等实体返回的结果进行安全评估后再使用。

此外，智能体作为被调用方还会感知到其他智能体调用请求，智能体作为被调用方措施：

- 严格校验调用方身份及权限范围，并进行访问控制；
- 对高频或异常调用请求实施速率限制和资源隔离；

- 记录所有调用请求日志并留存审计线索。

(2) 记忆安全

能够扫描判别记忆存储内容的安全性。能够保护记忆完整性，避免记忆被非法篡改。

(3) 规划安全

具备在面对不同输入或外部干扰时保持推理规划稳定。推理规划过程符合人类思维逻辑，依据突发事件和风险动态调整任务规划路径。提供规划决策过程的解释，对用户提供交互式解释支持。

(4) 行动安全

预先规定智能体行动执行范围，保持稳定性，防止因条件变化而出现不可预测的行为。对智能体任务的执行进度实时监控。

智能体作为调用方措施：

- 调用前验证被调用工具或智能体的身份合法性；
- 监控和记录调用工具的身份、权限和指令，用于后续安全审计。及时检测和隔离有故障的工具；
- 对敏感调用请求进行加密和完整性保护，防止中间人篡改；
- 遵循最小权限原则，仅向被调用方申请任务必需的权限。

(5) 交互安全

智能体模块之间，智能体与外界环境的通信应采用加密传输，避免数据在传输过程中被窃取，并定期更新加密密钥。定期核验和更新智能体的通信协议配置，防止因协议漏洞导致的攻击。确保智能体之

间，智能体与外部环境之间的交互符合预先定义的访问控制权限要求。配置智能体网络安全防护机制，阻止非法通信请求并拦截恶意流量，防止非法入侵行为。

智能体作为调用方措施：

- 对发出的请求进行数字签名，确保不可抵赖性；
- 敏感参数需脱敏处理后再传输。

数据安全管理贯穿智能体整个生命周期，具体措施如下：

具备数据完整性校验机制，检测数据在传输和处理过程中是否被篡改。保护用户隐私，对数据的敏感信息脱敏或者匿名化处理，防止个人信息泄露。保证使用的数据符合版权要求。建立最小授权的访问控制策略，确保智能体只能访问职责所需的最小必要的信息，且仅具备完成职责所需的最少的数据操作权限。具有数据备份机制。针对数据进行分类分级管理。

针对 3.2 的智能体安全风险列表，具体应对措施如下表所示。

表 8 智能体安全风险应对措施列表

威胁编号	应对措施
AIA01	输入过滤验证 提示词工程防护 模型安全对齐加固 恶意提示词检测
AIA02	数据加密传输 数据分级分类 数据脱敏、匿名化处理

威胁编号	应对措施
	差分隐私保护 细粒度动态访问控制 定期安全审计 数据备份
AIA03	第三方组件审查 模型完整性校验 安全开发流程 漏洞扫描更新 沙盒隔离运行
AIA04	细粒度动态访问控制 监控角色变更 敏感操作审计 跨代理权限隔离 双向身份认证 基于行为的身份伪造识别
AIA05	规划结果验证 决策边界管理 高风险操作人工确认 多源内容审核 决策过程透明可解释
AIA06	双向身份认证 数字签名 资源配额管控 系统负载监控

威胁编号	应对措施
	实时速率限制 智能体交互验证和监控 行为范围和自主性约束 系统干预机制 行为审计追溯 对关键决策过程实施多智能体共识验证
AIA07	安全通信协议 协议完整性核验和更新 定期协议审计 数字签名
AIA08	代码生成权限限制 沙盒环境隔离 脚本行为监控 执行控制策略 高权限代码人工审核
AIA09	全生命周期日志记录、监控和审计 日志加密签名
AIA10	记忆内容验证 会话隔离 异常检测系统 定期记忆清理 记忆快照取证
AIA11	细粒度动态访问控制 工具使用监控

威胁编号	应对措施
	指令过滤验证 行动范围约束 限制工具调用频次与范围 日志留存

以上措施对智能体的调用方和被调用方两种身份皆适用，其中：

1. 调用方更侧重：

- AIA04 对被调用实体进行身份认证；
- AIA07 对自身生成调用请求进行数字签名；
- AIA11 限制工具调用频次与范围；
- AIA02 对调用请求中敏感参数实施数据脱敏、匿名化处理。

2. 被调用方更侧重：

- AIA01 对调用请求做输入过滤验证；
- AIA04 对调用请求实施细粒度动态访问控制；
- AIA06 被调用过程中实施资源配额管控、系统负载监控和实时速率限制。

3.6 智能体安全风险标准治理应对措施

根据智能体安全风险经标准治理的改善程度，可将风险分为以下三种：

（1）可通过制定标准解决，风险具备明确的技术或流程边界，适合通过标准化手段进行系统性解决；其中风险 AIA03、AIA04、AIA07、AIA08、AIA11 属于该类风险，可以通过对智能体供应链、智能体交

互安全、智能体运行环境风险等相关领域进行标准化，制定技术规范、合规要求、接口协议等标准，来系统性解决该类型风险。

（2）需标准框架协同优化，风险无法被标准单独解决，但标准可为其建立评估框架、引领最佳实践，需与技术及监管协同应对；风险 AIA02、AIA06、AIA09 属于该类风险，可通过对人工智能数据生命周期安全，人工智能数据质量评估，多智能体协同安全，智能体日志审计等相关领域进行标准化，制定框架和实践参考，并配合算法改进，在系统创新等技术领域进步，促进该类型风险缓解。

（3）需标准提供基础支撑，风险根植于模型本质或属社会治理难题，其解决超越标准范围，主要依赖技术进步与法律监管，标准发挥对技术和监管的基础支撑作用；风险 AIA01、AIA05、AIA10 属于该类风险，可通过对智能体可信评估，智能体应用安全分类分级以及检索增强，多模态感知过滤等重要相关技术进行标准化，来支撑相关领域技术突破和监管措施。

4 智能体安全标准体系

4.1 智能体安全标准化需求

针对智能体新技术新应用带来的安全风险和挑战，结合智能体关键技术发展现状，从智能体的四个关键层面出发，提出以下五个方面的安全标准化需求。

(1) 通用共性需求

随着智能体技术飞速发展，各类技术框架层出不穷，应用场景推陈出新。国内外针对智能体的整体安全的威胁分析，安全需求框架等相关标准正在推进阶段。还应针对智能体安全风险开展应对措施研究，提出智能体安全框架和安全要求；应针对智能体各类工作模式和应用场景提出智能体分类分级安全方法，划分智能体应用安全等级，并提出各安全等级的安全管控措施和测评方法；智能体在各垂直行业中应用前景广阔，面临安全风险及技术瓶颈亟待优化、解决，因此应推出智能体应用场景安全实施指南，为智能体安全应用提供指引；应参照智能体安全技术要求，提出相应的智能体供应链、智能体安全测试方法，保障智能体在开发/引入、部署、运维全生命周期的安全。

(2) 感知层面需求

智能体已深度融合多模态技术，形成“感知→规划→记忆→行动”的通用架构框架。而多模态的感知融合，可能导致智能体从环境中获取的信息丰富、异构、难以审核或交叉过滤，从而给数据投毒、对抗

样本等攻击手段以可乘之机。因此，应针对感知模块获取过量、有毒信息等情况，开展感知信息的安全评估研究，提出安全评估指标，并实现数据清洗、过滤，安全存储，再执行后续操作。

(3) 规划层面需求

智能体在规划层面形成了多种规划模式，具备不同的自主程度和安全风险。例如，无固定工作流涉及多智能体的协作以及工作流全自主规划，可能导致规划结果不可控；端到端学习模式尚处在发展阶段，其训练过程不可控且不可解释，微调后的模型可控程度较低，人工审核和追溯也难以执行，安全风险较高。因此，应针对各等级自主程度的智能体实现安全分类分级治理，并开展智能体规划结果的安全评估。

(4) 记忆层面需求

智能体在记忆层面主要基于 RAG 技术，现有针对 RAG 技术的标准正在推进中。记忆技术对智能体性能提升显著，而智能体过多依赖记忆，也可能导致记忆泄露、记忆篡改、记忆投毒等攻击手段，对智能体以及智能体影响的环境和人带来巨大的安全风险。应开展智能体数据质量评估标准研究，分析、研判智能体数据质量，保障智能体记忆模块安全。此外，其他智能体记忆技术也尚在发展中，亟须对智能体记忆的格式、标签、内容审核和接口等领域进行统一标准化，实现智能体兼容多种记忆格式，推进智能体记忆新技术的有序发展。

(5) 行动层面需求

智能体在行动层面已存在 MCP/A2A 相关协议用于规范智能体与外界的交互格式和接口，国内外也推进多智能体通信、协作和互操作的标准化。但目前针对协议的安全威胁研究和安全要求以及安全措施的标准化研究还存在空白，亟须开展面向智能体行为过程中的交互、协同过程安全风险研究，并完成相应的互联安全标准研制，包括但不限于智能体交互安全、工具调用安全、协议安全、身份认证、运行状态认证、通信加密等重点安全问题。

4.2 智能体安全标准体系框架

结合现有人工智能安全标准体系，以及智能体安全标准化需求，构建适配智能体安全需求的智能体安全标准体系框架，如图 4。

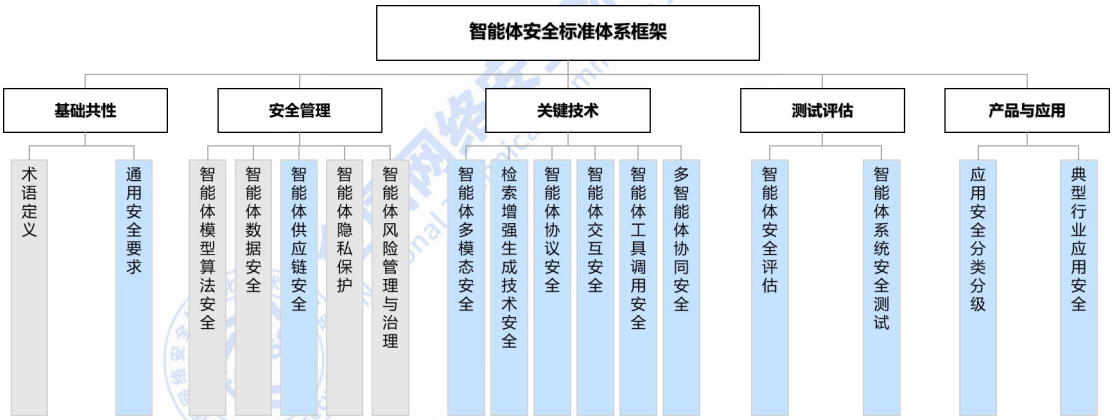

图 4 智能体安全标准体系框架

智能体安全标准体系框架包括 5 个安全维度，分别为基础共性、安全管理、关键技术、测试评估和产品与应用。

(1) 基础共性

基础共性部分，子维度包括术语定义和通用安全要求。其中，各维度安全标准主要内容如下：

- 术语定义：规范智能体及相关专用术语的概念和范围。

- 通用安全要求：包括智能体的安全框架和通用技术要求和模型优化等。

(2) 安全管理

安全管理部分，子维度包括数据安全、供应链安全、模型算法安全、隐私保护和风险管理及监管。其中，各维度安全标准主要内容如下：

- 数据安全：规范智能体数据安全要求，包括但不限于数据全生命周期安全要求、数据质量评估、内容合规检测和过滤等。
- 模型算法安全：规范智能体模型、算法在透明度、可解释性和价值对齐等方面的安全要求。
- 供应链安全：规范智能体在供应链安全方面的要求，包括但不限于智能体模型、算法来源可信、第三方库与插件安全、平台与运行环境安全等。
- 隐私保护：规范智能体在隐私方面的安全要求。
- 风险管理及监管：规范智能体的安全风险管理手段，以及对智能体的管理体系审核与认证机构要求、人类监管要求等。

(3) 关键技术

关键技术部分，子维度包括多模态安全、检索增强安全、协议安全、多智能体协同安全及工具调用安全。其中，各维度安全标准主要内容如下：

- 多模态安全：规范多模态智能体安全要求，包括但不限于多模

态对抗样本防护、跨模态一致性校验等。

- 检索增强生成技术安全：规范检索增强生成技术安全，包括但不限于检索增强生成技术全流程各环节的安全要求、应用安全要求等。
- 协议安全：规范智能体相关协议（如 A2A、ANP、MCP 等）的安全要求，包括但不限于身份认证安全要求、传输安全、访问控制等。
- 智能体交互安全：规范智能体交互安全要求，包括但不限于智能体交互过程和权限控制等。
- 工具调用安全：规范智能体调用相关工具的安全要求，包括但不限于智能体与被调用工具的身份互验、指令检验、权限控制等。
- 智能体协同安全：规范智能体协同安全要求，包括但不限于多智能体协同架构、身份管理、任务编排、指令校验等。

(4) 测试评估

测试评估部分，子维度包括安全评估和系统安全测试。其中，各维度安全标准主要内容如下：

- 安全评估：规范评估智能体决策结果和决策能力获取人类信任程度的关键评估指标，包括但不限于模型鲁棒性、公平性、隐私保护、可解释性、透明度、输出内容价值对齐、可靠性、合规性等。

- 系统安全测试：规范用于测试智能体性能是否符合技术要求的测试用例及评估方法。

(5) 产品与应用

产品与应用部分，子维度包括应用安全分类分级和垂直行业应用实施指南。其中，各维度安全标准主要内容如下：

- 应用安全分类分级：划分智能体应用安全等级，并提出各安全等级的安全管控措施。
- 典型行业应用安全：规范智能体在垂直行业安全应用的实施方法，涵盖软硬两方面智能体，并重点针对软智能体。

4.3 智能体标准演进分析

为系统构建智能体安全标准体系，本报告提出“演进与创新协同推进”的总体策略。在充分吸收现有通用安全标准框架的基础上，依托人工智能现有安全体系，进行系统性演进，遵循持续迭代、结构完善的建设路径。具体围绕以下三个核心维度展开标准布局，以实现体系化构建，避免资源重复投入，保障标准建设工作的系统性与可实施性：

(1) 基础共性维度：推进术语演进，夯实理论根基

在既有 GB/T 41867-2022《信息技术 人工智能 术语》体系与框架基础上，新增智能体相关术语的规范与定义。目前不具备智能体安全框架及要求，应针对智能体所具备的自主性、交互性与目标导向等新型技术特征，明确界定相关概念，为构建统一、科学的标准体系奠

定基础。

(2)安全管理维度：强化风险应对，优化安全监管

重点针对智能体特有的风险展开研究，对“风险管理与治理”相关标准进行演进，明确其在自主决策、环境交互等场景下的新型风险识别、评估与管控要求，以确保智能体风险管理与治理。

在算法安全层面，基于GB/T 46351-2025 《人工智能 多算法管理技术要求》、GB/T 42888-2023《信息安全技术 机器学习算法安全评估规范》、 GB/T 45225-2025《人工智能 深度学习算法评估》为核心依据，重点围绕智能体算法生命周期的统一管理、对抗样本的鲁棒性测试、模型公平性与偏见消减，以及输出内容的可控性等方面实现技术演进，确保智能体核心算法的可靠性、公正性与韧性。

在数据安全层面，基于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》、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》、GB/T 42755-2023《人工智能 面向机器学习的数据标注规程》、ISO/IEC 8183:2023《信息技术 人工智能 数据生命周期框架》等标准演进，保障智能体数据安全。

在隐私保护层面，基于GB/T 44588-2024 《数据安全技术 互联网平台及产品服务个人信息处理规则》进行演进，将隐私保护嵌入智能体设计与运行的全过程。

在风险监管与治理层面，以GB/T 46347-2025 《人工智能 风险

管理能力评估》为指引，着力于建立覆盖智能体全生命周期的风险动态识别机制、将安全评估流程嵌入研发关键环节、制定并演练应急响应预案，实现对智能体风险的常态化、体系化治理。

(3) 关键技术维度：填补空白维度，筑牢智能体安全防线

当前，智能体在关键技术领域仍面临安全挑战，亟待从以下维度构建和完善安全保障体系：

当前智能体供应链安全要求尚不完善，需建立智能体、模型、工具的供应链可信来源标准化验证机制，通过标准化数字签名、镜像完整性校验等措施保障安全，防止供应链投毒篡改。

智能体多模态安全技术要求有待明确，应针对多模态智能体有害内容交叉过滤、多模态一致性验证等制定标准化的安全要求。

智能体互联安全技术要求尚有待明确，应规定智能体与工具和其他智能体等外部实体的互联安全要求，包括加密、身份认证、权限管理，访问控制、完整性保护机制，防止中间人攻击、重放攻击、协议解析漏洞、工具滥用等风险。

多智能体协同安全还需标准化推进，需规定智能体间通信的内容过滤、冲突解决策略（如优先级仲裁、投票机制、资源竞争避免算法），作为多智能体系统协同的实践参考。

(4) 测试评估维度：聚焦行为验证，提升鲁棒性保障能力

在延续数据质量、模型性能等基础评估方法的同时，细化“系统测试”标准，覆盖智能体感知、决策、执行的完整闭环，保证智能体

在复杂应用场景中的鲁棒性、可控性与行为可靠性。

智能体安全评估需构建覆盖鲁棒性、可控性、符合性、透明性和可解释性等整体评估指标，需规定智能体可验证的输入恶意内容过滤，对输出内容的幻觉过滤能力等一系列指标，用于评估智能体防范提示词注入、越狱和幻觉等风险的能力。

智能体安全测试应制定针对智能体的测试标准，对智能体在四个能力层面上的安全指标以及不同场景的安全需求进行测试和评分。

(5)产品与应用维度：应用安全分类分级，形成多维保障框架

智能体应用安全分类分级尚缺乏针对性的标准化建设，需根据场景所属的垂直行业领域业务和安全需求，规范不同场景的安全分类分级，并给出适用于各个安全分级的智能体的安全要求。

智能体应用安全实施指南应结合行业最佳实践，为开发、部署与运维各阶段提供可操作的安全基线与落地路径，从而形成覆盖技术、管理与标准的多维保障框架。

5 智能体安全标准工作推进建议

随着人工智能发展进入“下半场”，智能体已成为推动产业变革的核心力量，但其新技术迭代与新应用落地也同步催生了数据泄露、决策失控、权限滥用等多元安全风险，对技术治理提出更高要求。智能体安全标准化是平衡技术创新与风险防控的核心支撑，需结合当前智能体感知、规划、记忆、行动等关键技术发展现状，依据《国家人工智能产业综合标准化体系建设指南（2024 版）》中“强化安全引领、推动标准落地”的核心要求与国际标准发展动态，构建“国内统一、国际协同、机制保障”的立体化推进体系。从以下三个方面提出工作推进建议。

5.1 构建智能体安全国家标准体系

紧扣智能体“感知→规划→记忆→行动”核心技术链路与全生命周期安全需求，构建覆盖基础共性、安全管理、关键技术、测试评估和产品与应用五个维度的智能体安全国家标准体系，为智能体安全治理提供统一技术基准与可操作、可落地的实施依据。

一是加快基础通用安全标准研制。聚焦智能体安全领域术语不统一、定级无依据、测评无规范等基础问题，**优先推进通用性、框架性、基础性标准研制，为后续技术标准与场景标准提供底层支撑。**建议近一两年制定《网络安全技术 智能体安全框架及要求》，明确通用智能体的安全基本要求，包括通用安全、各关键模块安全等要求。后续可继续制定《人工智能 智能体供应链安全技术要求》，规范智能体

在设计、开发、部署、运行等阶段引入数据、算法、插件等的安全要求。

二是制定安全关键技术标准。聚焦智能体“感知→规划→记忆→行动”4个关键模块，研制专项技术标准，**填补当前技术规范空白，阻断风险传导链条**。建议近一两年立项《网络安全技术 智能体互联安全要求》，规范智能体通过特定协议实现交互、调用工具过程中的安全要求，保障智能体行动安全。同时，建议制定《人工智能 多智能体协同安全基本要求》，基于交互、工具调用和协议等三个方面安全要求，规范智能体协同安全框架、协同调度安全、协同过程安全等要求，确保多智能体协同应用可落地、可推广。后续可深入研究智能体多模态安全影响因素及应对措施，制定《人工智能 智能体多模态安全技术要求》，规范多模态智能体有害内容过滤、多模态一致性验证等要求。

三是推进安全治理标准精准覆盖。**围绕智能体安全治理全流程需求**，建议近一两年制定《网络安全技术 智能体安全测试与评估方法》，并聚焦典型应用场景制定《人工智能 智能体应用安全分类分级方法》。前者提出智能体安全维度评估指标，指导实现对智能体安全维度进行评估量化，分析智能体安全程度，规范测试指标体系、测试方案编制、测试结论要求等；后者对智能体进行分类，并根据智能体应用在经济社会发展中的重要性和出现安全问题的影响程度，将智能体划分不同等级。为了应对典型应用场景的迫切需求，建议加快制定《网

络安全技术 智能体应用安全实施指南》系列，针对政务、具身智能、终端智能体、AI 浏览器、金融、工业制造、医疗等领域的差异化需求，提出针对性的智能体安全应用方法指南。根据指标制定的急迫性，可列出标准建议，如表 9。

表 9 智能体安全标准建议

时间	标准建议	关联风险	备注
近一两年开展	《网络安全技术 智能体安全框架及要求》	AIA01 至 AIA11	本标准旨在明确智能体各子系统及其能力应满足的安全要求，作为智能体安全标准体系的奠基性文件，为后续安全标准建设提供总体框架，为风险解决方增强智能体安全提供技术参考，并为行业及管理部门规范智能体产业与技术发展提供支撑。本标准提出智能体系统的整体安全框架，旨在明确智能体在感知、规划、记忆、行动等核心能力层面及交互层面的安全功能，界定智能体服务提供者、模型算法研发者与开发平台提供者在全生命周期中的安全职责要求，为相关方提供系统化的基础指引。
	《网络安全技术 智能体安全测试与评估方法》	AIA01 至 AIA11	本标准拟提出智能体安全测试与评估指标及评估方法。适用于指导智能体服务提供者、模型算法研发者、开发平台提供者、测评机构等开展智能体测评活动。
	《网络安全技术 智能体互联安全要求》	AIA04、AIA06、AIA07、AIA11	本标准拟规定智能体交互协议安全应用要求、如 A2A、ANP、MCP，包括但不限于身份认证安全、传输安全、访问控制等；交互过程安全和工具调用安全要求，如智能体互联过程的身份互验、指令检验、权限控制等。
	《网络安全技术 智	AIA01、	本标准确立了基于智能体应

时间	标准建议	关联风险	备注
	能体应用安全分类分级方法》	AIA05	用场景的垂直行业领域、业务及安全需求的安全分类分级方法，并规定了不同安全级别对应的安全要求。
	《网络安全技术 多智能体协同安全技术要求》	AIA01 至 AIA11	本标准分析提出智能体协同安全要求，旨在保障多智能体协同安全，降低潜在安全风险，为形成安全可信互联互通的 AI 智能体生态体系奠定基础。
	《网络安全技术 智能体应用安全实施指南》系列	AIA01 至 AIA11	本系列指南针对政务、具身智能、终端智能体、AI 浏览器、金融、医疗、工业制造、教育等典型应用场景，提出智能体安全实施的系统性方法与操作指南，为智能体在各行业的落地提供可操作的安全实践依据，降低因场景差异性引发风险。
后续开展	《网络安全技术 智能体供应链安全技术要求》	AIA03、AIA04、AIA08、AIA09	本标准规范了智能体在开发、集成、部署与运维过程中所涉及的数据、模型、插件、平台及运行环境等供应链环节的安全要求，确保智能体从源头到部署全过程的可信可控。
	《网络安全技术 智能体多模态安全技术要求》	AIA01 至 AIA11	本标准规定了基于大模型的多模态智能体在感知、规划、记忆、行动等环节应满足的安全技术要求，重点包括多模态输入内容的安全过滤、跨模态一致性校验、对抗样本防护、生成内容真实性保障、敏感信息防泄露、记忆系统防篡改等技术规范。

5.2 深化国际标准协同与对接机制

以“主动参与、优势输出、互认互通”为原则，深度融入全球智能体安全标准治理体系，提升国际话语权。

一是精准对接核心国际标准化组织。针对国际标准化组织（ISO）、

国际电工委员会（IEC）、国际电信联盟（ITU）等机构的智能体安全标准工作重点，构建分层对接体系，输出中国实践经验，争取核心指标的主导权。

二是牵头制定特色领域国际标准。依托我国在消费级智能体、多智能体协作等领域的产业优势，发起国际标准提案。例如在智慧文旅、城市服务等场景，推动制定消费级智能体安全交互相关标准，规范多模态输入过滤与隐私保护要求；在工业互联网领域，研制多智能体协同安全协议相关标准，解决跨系统交互中的信任认证与风险隔离问题。

三是推动国内标准与国际标准互认机制。选取成熟国家标准，与核心国际标准比对分析，形成差异对照表与适配指南；联合第三方检测机构建立国际互认的测试平台，实现一次检测、多标准认证。针对跨境智能体应用，探索“标准互认+合规声明”的监管模式，降低企业国际化合规成本。

5.3 完善标准实施与支撑保障机制

构建“政府引导、市场主导、多方参与”的工作体系，确保标准从研制到落地的全链条实效。

一是健全标准化工作协调机制。依托网络安全标准化技术委员会（TC260），统筹国家标准、行业标准、团体标准制订计划，建立“基础共性国标先行、技术创新团标突破、场景需求行标细化”的协同机制。建立覆盖企业、高校、检测机构的多元化起草团队，确保标准技

术先进性与产业适用性。

二是强化标准测试验证能力。依托工业和信息化部电子第五研究所、电子标准化研究院、国家工业信息安全发展研究中心等机构，建设国家级智能体安全标准验证平台，配备提示词注入、工具滥用等专项测试环境，提供标准符合性检测服务。推动测试平台与上海人工智能实验室、中关村实验室、清华大学等单位联动，开展标准预研与技术验证，实现“标准研制—测试反馈—迭代完善”的闭环管理。

三是构建标准推广应用生态。在“人工智能+”行动重点领域，选取智慧医疗、先进制造等场景开展标准试点，总结可复制的应用模式；将标准符合性纳入“专精特新”企业认定、科技项目申报的评价指标，引导企业主动应用标准；建立标准信息公共服务平台，提供标准解读、案例分享、合规咨询等一站式服务。

6 总结与展望

智能体正经历从“高级工具”到“自主伙伴”的范式转变，成为人工智能变革的核心驱动力。其发展呈现三大趋势：一是技术深度融合，智能体正从单一语言模型向融合视觉、听觉、环境感知的多模态智能演进；二是自主性显著增强，智能体实现了目标自动分解与动态规划，从依赖专家预设的“固定工作流”向“无固定工作流”的自主规划演进，显著提升了在非结构化环境中的问题解决能力；三是社会角色深化，产业界对 A2A、MCP 等互操作协议的推动，正打破智能体间的数据与交互壁垒，为构建多智能体协作系统、形成社会性智能奠定基础。

这种技术演进将引发社会形态的深刻变革。在经济领域，智能体不仅是生产力工具，更可能成为独立的经济主体，参与市场交易、签订智能合约甚至创造经济价值。法律体系面临重构压力，需要确立智能体的法律地位、权利义务边界和责任认定机制。当智能体能够自主完成价值创造时，传统以自然人和法人为基础的法律框架将面临挑战，可能需要创设“电子人格”等新的法律概念。社会治理层面，智能体可提升公共服务效率，实现城市管理的精准化与应急响应的智能化。但另一方面，智能体系统的复杂性可能导致新型社会风险。例如，金融市场中智能体的同质化决策可能放大系统性风险；医疗、交通等关键领域智能体的集中失效可能引发社会危机；智能体之间的非透明

协作可能形成难以监管的“算法共谋”。

我国推动智能体发展，开展智能体安全治理，应注重以下三点：

1、注重平衡安全与发展、创新与规范的关系。

既要鼓励技术创新，保持产业竞争力，也要建立符合国情的安全标准与伦理框架，防范技术滥用带来的社会风险。特别是在基础模型研发、测试验证平台建设等关键环节，需要发挥新型举国体制优势，实现重点突破。

2、建立适应智能体特性的新型监管范式。

首先，应建立智能体全生命周期监管框架，从研发备案、运行监测到失效追溯形成完整管理闭环。其次，需推进监管科技创新，发展实时监测智能体行为的技术，实现动态风险预警。特别是在国防、金融等关键领域，可能需要设立智能体运行“黑匣子”制度，确保行为可追溯、决策可审计。

3、智能体安全标准化工作需进行前瞻性布局与深化。

持续明确智能体在关键领域的辅助型定位，确保人类主导。首先，通过分类分级的精准规制，为不同风险等级的应用场景设定差异化的合规要求。其次，建立标准与技术的动态协同更新机制，快速响应技术迭代。积极参与并主导国际标准制定，推动形成全球互认的安全认证体系。最后，随着多智能体协作成为常态，标准重点应从单个智能体安全转向多智能体系统的协同安全。

智能体的未来充满无限潜力，但其健康发展必须建立在坚实的安

全地基之上。通过构建一个科学、前瞻、开放的安全标准体系，并实施与之配套的敏捷治理策略，确保智能体技术真正成为造福社会、提升人类福祉的可靠力量，最终实现安全与发展并重的长远目标。

附录 A 相关智能体术语

1.智能体 AI agent （AIA）

automated entity that senses and responds to its environment and takes actions to achieve its goals.

automated is "pertaining to a process or system that, under specified conditions, functions without human intervention".

泛指一种自动化实体，即在指定条件下无需人工干预或受控人工干预即可运行，能够感知并响应其环境，并采取行动以实现其目标的流程或系统。

注：此处的“自动化”特指“在特定条件下，流程或系统的运作无需人工干预”。

[来源：ISO/IEC 22989:2022[34] 有修改]

2.感知和认知 Perception and cognition

The perception and cognition comprise of recognition, comprehension, dialogue, generation, and reasoning.

感知和认知包括识别、理解、对话、生成和推理。

[来源：ITU-T F.748.46[35]]

3.规划 Planning

The planning includes task planning, task scheduling and task optimization.

规划包括任务规划、任务调度和任务优化。

[来源: ITU-T F.748.46]

4.记忆 Memory

The memory includes short-term memory and long-term memory.

记忆包括短期记忆和长期记忆。

[来源: ITU-T F.748.46]

5.行动 Action

The action includes virtual environment execution capability and real-world environment execution capability.

行动包括虚拟环境执行能力和现实世界环境执行能力。

[来源: ITU-T F.748.46]

6.模型上下文协议 Model Context Protocol（MCP）

The Model Context Protocol （MCP） is an open protocol that enables seamless integration between LLM applications and external data sources and tools.

模型上下文协议（MCP）是一项开放协议，可实现大型语言模型（LLM）应用与外部数据源及工具之间的无缝集成。

[来源: The official specification, documentation and implementations for the Model Context Protocol （MCP） [36]]

7.智能体对智能体协议 Announcing the Agent2Agent

Protocol （A2A）

The Agent2Agent （A2A） Protocol is an open standard designed to enable seamless communication and collaboration between AI agents. In a world where agents are built using diverse frameworks and by different vendors, A2A provides a common language, breaking down silos and fostering interoperability.

Agent2Agent（A2A）协议是一项开放标准，旨在实现人工智能代理之间的无缝通信与协作。在当前代理由不同厂商采用多种框架构建的环境下，A2A 提供了一种通用语言，打破了信息孤岛，促进了互操作性。

[来源: Announcing the Agent2Agent Protocol （A2A）[37]]

8.智能体网络协议 Agent Network Protocol（ANP）

Agent Network Protocol（ANP）提出了面向智能体互联网的新一代通信协议。ANP 坚持 AI 原生设计，兼容并复用现有互联网协议，采用模块化可组合架构，遵循极简而可扩展的原则，并基于现有基础设施实现快速部署。通过三层协议体系——身份与加密通信层、元协议协商层、应用协议层，ANP 系统性地解决了智能体身份认证、动态协商及能力发现互操作问题。

[来源: Agent Network Protocol 技术白皮书 [38]]

9.工具 tool

智能体可能调用的功能实体，包含且不限于网络服务、应用程序、物理环境中存在的感知实体或操作实体等实体。

附录　B 国内外已发布及在研相关标准

本附录收录与智能体安全密切相关的标准和研究报告。为构建完整的技术参考体系，聚焦智能体安全核心标准的同时，也扩展收录了在基础框架、技术要素层面关联紧密的人工智能通用安全标准及智能体标准，以提供更全面的标准化现状视图。

B.1 已发布国际标准

标准号	标准名	归口单位
ISO/IEC 22989:2022	Information technology — Artificial intelligence — Artificial intelligence concepts and terminology 信息技术 人工智能 人工智能概念和术语	ISO/IEC JTC1/SC42
ISO/IEC 23053:2022	Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML) 使用机器学习的人工智能系统框架	ISO/IEC JTC1/SC42
ISO/IEC 24029-1:2021	Artificial Intelligence (AI) — Assessment of the robustness of neural networks — Part 1: Overview 人工智能 神经网络鲁棒性评估 第1部分：概述	ISO/IEC JTC1/SC42
ISO/IEC TR 27563:2023	Security and privacy best practices for AI use cases AI用例中的安全和隐私最佳实践	ISO/IEC JTC1/SC27
ITU-T F.748.46（ex F.TE-AIA）	Requirements and evaluation methods of artificial intelligence agents based on large scale pre-trained model	ITU

B.2 已发布国内标准

标准号	标准名	归口单位
GB 45438-2025	网络安全技术 人工智能生成合成内容标识方法	TC260
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	TC260
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	TC260
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	TC260

TC260-PG-2025NA	网络安全标准实践指南——人工智能生成合成内容标识 服务提供者编码规则	TC260
GB/T 42888-2023	信息安全技术 机器学习算法安全评估规范	TC260
GB/T 45958-2025	网络安全技术 人工智能计算平台安全框架	TC260
T/CCSA 800-2026	人工智能 智能体系统能力分级方法和技术要求	CCSA

B.3 在研国际标准

标准号	标准名	归口单位
ISO/IEC 27090	Cybersecurity — Artificial intelligence — Guidelines for addressing security threats and failures in AI systems 网络安全 人工智能解决人工智能系统中安全威胁和故障的指南	ISO/IEC JTC1/SC27
ISO/IEC 27091	Cybersecurity and Privacy — Artificial Intelligence — Privacy protection 网络安全与隐私-人工智能-隐私保护	ISO/IEC JTC1/SC27
ISO/IEC PWI 25240	Evaluation of AI-based Technology 基于人工智能技术的评估	ISO/IEC JTC1/SC27
ITU-T X.S-A1A	Security Requirements and Guidelines for Artificial Intelligence Agent 智能体安全要求与指南	ITU
ITU-T X.sr-ai	Security requirements for AI systems 人工智能系统安全要求	ITU
ITU-T X.sgGenAI	Security Guidelines for Generative Artificial Intelligence Application Service 生成式人工智能应用服务安全指南	ITU
ITU-T TR.se-ai	Security Evaluation on Artificial Intelligence technology in ICT 信息通信技术中人工智能技术的安全评估	ITU
ITU-T F.AIAP	Framework and requirements for AI agents platform 人工智能代理平台框架与要求	ITU
ITU-T F.RF-AIAC-FM	Requirements and Framework for AI Agent Collaboration for Foundation Models 基于基础模型的人工智能代理协作要求与框架	ITU
ITU-T X.sr-taimas	Security requirements for terminal-based artificial intelligence multi-agent system	ITU

	基于终端的人工智能多代理系统安全要求	
ITU-T X.gavd-mas	Guidelines for application vulnerability detection based on multi-agent system 基于多代理系统的应用漏洞检测指南	ITU
ITU-T TR.sem-AIA	Security evaluation methods for an artificial intelligence agent 智能体安全评估方法	ITU
ITU-T B.agai	Design Principles for Identity Management (IM) in the Context of Agentic AI 智能体背景下的身份管理设计原则	ITU
ITU-T TR.ltf-AAI	Landscape of Trust Framework for Agentic AI 智能体 AI 信任框架全景图	ITU

B.4 在研国内标准

标准名	归口单位
网络安全技术 人工智能模型即服务安全要求	TC260
网络安全技术 人工智能应用安全运营能力评估规范	TC260
人工智能 智能体互联 第 1 部分：总体架构	TC28/SC42
人工智能 智能体互联 第 2 部分：身份码	TC28/SC42
人工智能 智能体互联 第 3 部分：身份管理	TC28/SC42
人工智能 智能体互联 第 4 部分：智能体描述	TC28/SC42
人工智能 智能体互联 第 5 部分：智能体发现	TC28/SC42
人工智能 智能体互联 第 6 部分：智能体交互	TC28/SC42
人工智能 智能体互联 第 7 部分：智能体工具调用	TC28/SC42
人工智能 智能体平台通用技术要求	TC28/SC42
人工智能 关键技术 智能体技术规范	TC28/SC42
人工智能 电力智能体通用技术要求	TC28/SC42
人工智能 多模态智能体技术要求	TC28/SC42
AI 智能体安全 第 1 部分 基本框架和安全要求	CCSA
人工智能关键技术 智能体基础技术能力要求	CCSA
生成式人工智能终端 智能体技术要求和测评方法	CCSA
人工智能安全治理 生成式人工智能图像检测服务系统能力要求	CCSA
基于人工智能的图像识别服务鲁棒性安全要求和测试方法	CCSA
基于人工智能的音视频合成服务安全管理要求	CCSA
基于人工智能的音视频合成服务安全评估要求	CCSA
电信网和互联网安全大模型测评指标及方法 网络安全领域	CCSA
电信网和互联网安全大模型能力要求 总体框架	CCSA
电信网和互联网安全大模型能力要求 网络安全领域	CCSA
电信网和互联网安全大模型应用架构	CCSA
电信和互联网人工智能算法安全要求	CCSA

电信网和互联网安全大模型能力要求 数据安全领域	CCSA
基于大模型的网络安全威胁信息检索技术要求	CCSA
人工智能数据安全保护框架	CCSA
电信和互联网人工智能数据安全评估方法	CCSA
人工智能数据安全通用要求	CCSA
人工智能服务平台数据安全要求和评估方法	CCSA
大模型检索增强生成技术和应用安全要求	CCSA
检索增强生成（RAG）能力技术要求	CCSA TC601
大模型检索增强生成技术和应用的要求和评估方法	CCSA

附录 C 智能体分类与典型产品实例对照表

本附录旨在将本报告第 1.2 节“智能体分类与典型应用”中提出的理论分类，与当前国内外市场上具有影响力的智能体产品或项目进行对应，以提供更直观、具体的参考。

分类维度	主要类型	典型产品/项目实例
按产品形态划分	软智能体	国际：ChatGPT、Claude、GitHub Copilot、AutoGPT、Microsoft 365 Copilot、Google Gemini Advanced、Devin（Cognition AI，首个自主编程智能体）、Grok（xAI）、Manus（Butterfly Effect，已被 Meta 收购） 国内：豆包（字节跳动）、文心一言（百度）、通义千问（阿里云）、Kimi Chat（月之暗面）、智谱清言（智谱 AI）、混元（腾讯）、星火（讯飞）、扣子/Coze（字节跳动，Agent 开发平台）、AutoGLM（智谱 AI，手机操作 Agent）、跃问（阶跃星辰）、元宝（腾讯）
	硬智能体	国际：Tesla Autopilot/FSD（特斯拉，自动驾驶）、Boston Dynamics Atlas/Spot（波士顿动力，机器人）、DJI 无人机（大疆，智能飞行系统） 国内：小米 CyberDog（机器人）、百度 Apollo（自动驾驶）、蔚来 NOP（自动驾驶）、小鹏 XNGP、华为 ADS、大疆车载
按功能与场景划分	智能助理/个人助手	国际：Apple Siri、Google Assistant、Amazon Alexa、Grok（xAI）、Perplexity（AI 搜索引擎） 国内：小爱同学（小米）、小度（百度）、天猫精灵（阿里）、豆包手机助手（字节跳动）、YOYO（荣耀）、小艺（华为）、海螺 AI（MiniMax）、万知（零一万物）
	代码生成/开发助手	国际：GitHub Copilot、Tabnine、CodeWhisperer（AWS）、Claude、Devin（Cognition AI，自主编程 Agent）、Cursor、Windsurf 国内：阿里云通义灵码、百度 Comate、

分类维度	主要类型	典型产品/项目实例
		腾讯 AI 编程助手、CodeGeeX(智谱 AI)、Fitten Code（Fittentech）
	企业/办公助手	国际：Microsoft 365 Copilot、Salesforce Einstein、ServiceNow AI 国内： 钉钉 AI 助理、飞书智能伙伴、WPS AI、如流（百度）
	内容创作/生成	国际：Midjourney、Sora、Runway 国内：剪映 AI、快影 AI、可灵 AI（快手）
	搜索/研究助手	国际：Perplexity、Grok DeepSearch、ChatGPT Search 国内：秘塔 AI 搜索、360AI 搜索、天工 AI 搜索
	手机/设备操作	国际：Apple Intelligence（苹果，操作系统级智能体）、Operator（OpenAI，设备操作框架） 国内：AutoGLM（智谱 AI，手机 Agent）、豆包手机助手、YOYO（荣耀）
	通用任务/自主执行	国际:Manus(通用自主 Agent)、AutoGPT 国内：扣子空间（字节）、AutoGLM 沉思（智谱 AI）
按数量与协作划分	单智能体系统	绝大多数上述“软智能体”个人应用、如 ChatGPT 基础对话、单个客服机器人。
	多智能体系统	国际：Meta 的 CICERO（游戏外交）、AutoGen（微软）、Multi-Agent Orchestration（LangGraph）、CrewAI、Microsoft Magentic-One 国内： 阶跃星辰的“跃问”、 扣子 Coze 多 Agent 模式（字节跳动）、AutoGen 中文社区
	人机协作	主要应用于工业和医学领域: 工业：AI 视觉检测的工人辅助系统（腾讯云工业质检）、预测性维护系统（西门子 MindSphere）、协作机器人（ABB YuMi） 医疗：AI 辅助诊断系统（腾讯觅影）、手术机器人（达芬奇手术机器人）
按决策依据划分	确定性智能体	基于严格规则引擎的聊天机器人、早期的工业流水线控制程序、RPA（机器人

分类维度	主要类型	典型产品/项目实例
		流程自动化）、IFTTT
	非确定性智能体	所有基于大语言模型（LLM）驱动的智能体，包括 GPT 系列、Claude、Gemini、文心一言等所有大模型驱动 Agent
按决策过程划分	简单反射智能体	智能家居传感器（如人体感应开灯）、早期关键词匹配客服机器人、最简单的 Chatbot
	基于模型的反射智能体	传统自动驾驶（维护车辆状态模型）、Siri/小爱同学（结合用户历史偏好）、银行 APP 智能客服（记住办理进度）
	基于目标的智能体	导航软件（路径规划）、AutoGPT（目标分解执行）、Devin（完成指定编程任务）、Manus（完成用户指定目标）
	基于效用的智能体	智能投顾（权衡收益风险流动性）、推荐系统（多目标优化）、自动驾驶决策（安全 vs 效率权衡）
	学习型智能体	国际：DeepMind AlphaGo/AlphaFold、ChatGPT（RLHF 学习）、Devin（从错误中学习） 国内：AutoGLM（智谱 AI，自演化学习框架）
按所处层次划分	操作系统智能体	国际：Apple Intelligence（深度集成于 iOS/macOS）、Google Pixel AI（谷歌）、Samsung Galaxy AI（三星） 国内：HyperMind（小米）、蓝心小 V（vivo，接入 OriginOS）、MagicOS（荣耀，YOYO 智能体）、鸿蒙 Next（华为，小艺）、ColorOS（OPPO，小布）
	应用智能体	绝大多数以独立 App 形式存在或为特定软件（如 Office、Photoshop）提供增强功能的 AI 插件/助手。 • 浏览器插件：有道灵动翻译（网易）、Copilot（微软）、Monica（Monica） • 办公软件插件：Office Copilot（微软）、WPS AI（金山） • 专业工具插件：Photoshop AI（Adobe）、Neural CAD（Autodesk）

参考文献

[1] 《AI Agents Market by Agent Role （Productivity & Personal Assistants, Sales, Marketing, Customer Service, Code Generation）, Agent Systems （Single Agent, Multi Agent）, Product Type （Ready to Deploy Agents, Build Your Own Agents）－Global Forecast to 2030》[R]. Markets and Markets, 2024.
[2] 斯坦福大学. 《人工智能指数报告（AI Index Report 2025）》[R]. 2025.
[3] 美国政府.《关于安全、可靠、值得信赖地开发和使用 AI 的行政命令》[Z]. 2023.
[4] 欧盟.《人工智能法案》[Z]. 2024.
[5] 欧盟.《人工智能责任指令》[Z]. 在研.
[6] 加拿大政府.《人工智能和数据法》（AIDA）[Z]. 在研.
[7] 加拿大政府.《生成式人工智能行为守则》[Z]. 2023.
[8] 新加坡政府.《用于生成式人工智能的人工智能模型管理框架》[Z]. 2024.
[9] 韩国政府.《AI 框架法案》[Z]. 2024.
[10] 国务院.《新一代人工智能发展规划》[Z]. 2017.
[11] 全国人大常委会.《中华人民共和国网络安全法》[Z]. 2016.
[12] 全国人大常委会.《中华人民共和国数据安全法》[Z]. 2021.
[13] 全国人大常委会.《中华人民共和国个人信息保护法》[Z]. 2021.
[14] 国家互联网信息办公室.《互联网信息服务算法推荐管理规定》[Z]. 2021.
[15] 国家互联网信息办公室， 工业和信息化部， 公安部.《互联网信息服务深度合成管理规定》[Z]. 2022.
[16] 国家新一代人工智能治理专业委员会.《新一代人工智能伦理规范》[Z]. 2021.
[17] 国家互联网信息办公室， 等.《生成式人工智能服务管理暂行办法》[Z]. 2023.
[18] 国务院.《网络数据安全管理条例》[Z]. 2024.
[19] 国务院.《关于深入实施“人工智能+”行动的意见》[Z]. 2025.
[20] 国家互联网信息办公室， 等.《人工智能安全治理框架 2.0》

[Z]. 2025.
[21] 许晓耕、高超、郝春亮. 《人工智能安全标准化最新进展》[J]. 信息技术与标准化, 2025（04）:69-75.
[22] 蔡亚森. 《中国人工智能标准发展趋势》[C]//中国标准化协会. 中国标准化年度优秀论文（2022）论文集. 2022.
[23] 张世天、范博、郝春亮. 《人工智能安全新进展标准化研究》[J]. 信息技术与标准化, 2023（9）:11-15.
[24] 全国网络安全标准化技术委员会.《人工智能安全标准体系（V1.0）》（征求意见稿）[Z].2025.
[25] 中国电子技术标准化研究院, 清华大学, 百度, 华为, 360, 阿里巴巴, 等.《人工智能安全标准化白皮书(2019版)》[R/OL]. 2019.
https://www.tc260.org.cn/portal/article/1/20191031151659.
[26] 全国网络安全标准化技术委员会.《人工智能安全标准化白皮书（2023版）》[R/OL]. 2023.
https://www.tc260.org.cn/portal/article/2/20230531105159.
[27] 中国电子技术标准化研究院, 等.《人工智能标准化白皮书（2018版）》[R/OL]. 2018.
https://www.cesi.cn/201801/3545.html.
[28] 中国电子技术标准化研究院, 等.《人工智能标准化白皮书（2021版）》[R/OL]. 2021.
https://www.cesi.cn/202107/7796.html.
[29] 国家人工智能标准化总体组.《人工智能伦理风险分析报告》[R/OL]. 2019.
https://www.cesi.cn/images/editor/20190425/20190425142632634001.pdf.
[30] 国家人工智能标准化总体组.《人工智能开源与标准化研究报告》[R/OL]. 2019.
https://www.cesi.cn/images/editor/20190425/20190425143108915001.pdf.
[31] OWASP.《OWASP Top 10 for LLM Apps & Gen AI Agentic Security Initiative》[R/OL]. 2025.
[32] 上海人工智能实验室.《终端智能体安全2025》[R]. 2025.
[33] 360, 清华大学.《智能体安全实践报告》[R]. 2025.

[34] International Organization for Standardization/International Electrotechnical Commission.《Information technology - Artificial intelligence - Artificial intelligence concepts and terminology: ISO/IEC 22989:2022》[S]. 2022.
[35] International Telecommunication Union Telecommunication Standardization Sector.《Functional framework and capabilities of artificial intelligence enabled networking: ITU-T F.748.46》[S]. 2023.
[36] Model Context Protocol.《The official specification, documentation and implementations for the Model Context Protocol (MCP) [EB/OL]》. 2025. https://modelcontextprotocol.io/.
[37] Google.《Announcing the Agent2Agent Protocol (A2A) [EB/OL]》. 2025. https://developers.googleblog.com/en/a2a-a-new-era-of-agent-interoperability/.
[38]《Agent Network Protocol 技术白皮书》[R]. ANP 社区, 2024.