CAICT 中国信通院

集智 专题报告

智能体治理研究报告

（2026 年）

中国信息通信研究院政策与经济研究所

中国政法大学人工智能法研究院

2026年9月

版权声明

前　言

2026 年，全球智能体技术加速成熟，产业应用爆发增长。智能体凭借自主感知、规划决策与工具调用能力，推动人工智能实现从“能说话”向“能办事”的关键跨越，人机交互向主动化演进，网络形态向“万智互联”重构，经济运行向智能经济转型，人工智能革命的图景日益清晰。与此同时，智能体自主性、交互性与学习性特征叠加放大，推动风险从输出偏差向行动失控、从单点故障向网络传导、从静态缺陷向动态变异演进，对现有治理体系构成了前所未有的全新挑战。

面对智能体时代的治理新命题，人工智能治理已全面进入从以内容合规、事前审查为重心的模型治理，向以行为约束、全周期治理为重心的智能体治理加速转型的关键阶段。主要经济体纷纷围绕权限管控、风险分级、生态优化等方面细化规则标准，力图在促进规模化应用与防范系统性风险之间寻求动态平衡。报告立足 2026 年智能体技术演进与全球治理实践的最新动态，系统梳理智能体时代的新形势新变革与 AI 治理新命题，聚焦身份可信、权限可控、数据可管的基础层，审查评估、风险分级、透明披露、风险干预的运行层，责任配置、可信互联、竞争有序的生态层，构建智能体三层治理框架，并从理念引领、制度完善、技术创新、生态构建四个维度，对智能体治理未来提出展望。希望研究成果能够为社会各界准确把握智能体治理趋势、积极参与智能体治理实践提供有益参考。

目 录

图 目 录

表 目 录

一、智能体时代新形势新变革

2026 年，智能体加速从实验室走向规模化应用，从被动应答的“对话框”演变为能够理解目标、拆解任务、调用工具的数字“代理人”。总体来看，智能体能力跃升但仍存在明显短板，市场竞争格局加速形成，部署形态日益分化；与此同时，智能体带来人机交互、网络形态与经济运行的系统性重构。准确把握智能体的技术能力、产业动态与范式影响，是理解智能体治理新命题的前提。

（一）智能体从感知环境向行动自主跨越

1.内涵特征：从被动应答的“对话框”到自主行动的数字“代理人”

智能体是指能够感知环境、推理决策、执行动作以实现特定目标的智能计算系统，具有自主性、交互性、反应性和适应性等特征，体现了人工智能从“对话式工具”向“行动式代理人”的形态跃迁。从系统架构看，智能体主要由感知、记忆、规划、行动四大模块构成。**感知模块是智能体的“输入接口”，**负责接收和解析外部环境和设备中多源异构信息，涵盖文本、图像、音频、视频、传感器数据等；**记忆模块是智能体的“经验库”，**由短期记忆和长期记忆构成，前者存储当前任务相关的上下文信息，协助规划模块精准判断并维持任务执行的一致性；后者持久化保存案例、历史经验及用户基本信息等跨任务内容，赋予智能体持续学习与个性化服务能力；**规划模块是由大模型驱动的认知和调度中枢**，基于感知输入和记忆内容进行分析推理，生成目标导向的策略或行动计划；**行动模块则是智能体的执行终端，**

将规划决策转化为作用于外部环境的具体操作——对数字环境主要依托 GUI、API、CLI 等交互通道调用工具与外部资源，对物理环境则依托机器人等具身载体完成操作。在上述模块化架构之上，智能体还需依托开放协议构建生态接口，MCP 等协议解决模型与工具、数据源的标准化连接，A2A 等协议解决智能体之间的通信与协作，二者共同支撑人机交互、跨应用操作与多智能体协同的实现。

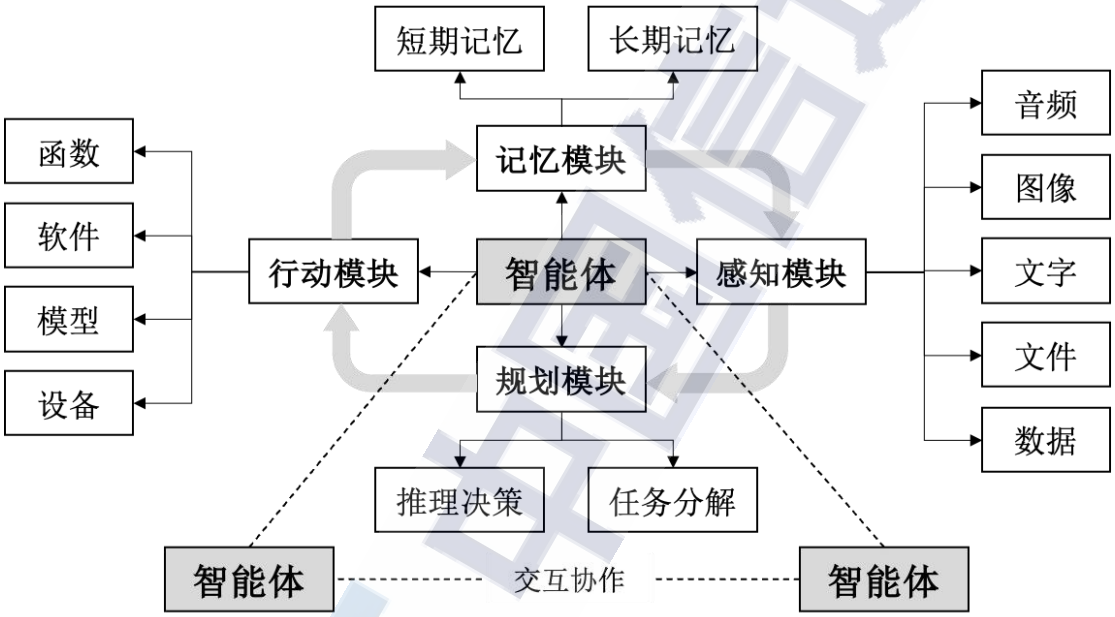

图 1 智能体技术架构

2.能力现状：基准测试成绩跃升，长程复杂任务仍存明显短板

当前，智能体能力正沿“准确性提升”与“任务时长拓展”两条主线快速跃升，但呈现明显的“锯齿状”特征，即短链条任务已接近甚至超越人类水平，长程复杂任务的成功率却断崖式下跌。**一是通用能力大幅跃升，与人类基准的差距快速收窄。**根据 GAIA 通用能力测试数据，智能体准确率已从 2025 年 1 月的 30%提升至 9 月的 74.5%，

与人类基准的差距收窄至约 17.5 个百分点[1]；截至 2026 年 6 月，CustomGPT.ai 以 93.36%的综合得分，首次超越人类平均基准水平（约 92%），标志着标准化通用任务上，头部智能体已整体跨越人类平均线。**二是可持续执行的任务时长呈现指数级拓展。**METR 时间跨度指标显示，2026 年初智能体以 50%可靠性完成的任务时长已接近 12 小时，且该指标约每 7 个月翻一番[2]。若这一趋势延续，智能体有望在数年内具备连续执行“周级”乃至“月级”任务的潜力。**三是长程复杂任务仍是突出短板，能力水平随任务链条拉长而衰减。**METR 同期数据显示，超过 4 小时人类工作时长的任务成功率不足 10%；OSWorld2.0 测试进一步显示，对人类预估耗时超过 163 分钟的任务，智能体完成率几乎全部归零[3]。究其根源，智能体单步误差会沿任务链条逐级累积放大；同时，长程任务对记忆持久性和规划稳定性提出更高要求，而当前智能体普遍存在长上下文记忆衰减、执行中目标漂移、错误自纠能力不足等短板，导致任务步数越多、周期越长，失败概率越高。

综合来看，智能体能力跃升与短板并存，在边界清晰、步骤较少的短链条任务上已具备实用价值，但在多步骤、长周期、强协同的复杂任务上可靠性仍然不足，“能用”与“好用”之间仍有距离。

3.演进阶段：正从单一智能体走向协作代理生态系统

从技术迭代的视角看，Gartner 将智能体的发展划分为五阶段成

[1] 参见 GAIA 通用人工智能助手基准测试排行榜（Hugging Face），2026 年。
[2] METR，“Time Horizon 1.1”，2026；arXiv:2602.02289。
[3] 参见 OSWorld 2.0 基准测试，2026 年。

熟度模型，系统勾勒出智能体从简单辅助工具演进为高度自主智能体生态的完整路径。**第一阶段是嵌入式助手**（2025 年）。AI 能力以“功能插件”的形式嵌入现有软件与应用之中，尚不具备独立产品形态，工作方式以被动响应为主，依赖用户主动发起指令，完成信息检索、内容生成、简单问答等任务，典型形态如各类办公软件中的 Copilot 类功能。**第二阶段是任务特定代理**（2026 年）。智能体开始针对单一、明确的任务场景进行专门设计，具备从任务规划到落地执行的完整闭环能力，可在有限人工监督下独立完成特定任务，如自动编写代码、生成数据分析报告、处理客服工单等。与第一阶段相比，其核心跃迁在于从“被动应答”转向“主动执行”。**第三阶段是协作代理生态系统**（2028—2029 年）。多个具备不同专长的专用代理被组合起来，通过任务分解、分工协作与相互调用，共同完成跨职能、跨流程的复杂任务，形成类似人类组织中的“代理团队”。例如，市场分析、文案生成、投放执行等多个代理可协同完成一次完整的营销活动。**第四阶段是自主代理网络**（2028—2030 年）。代理具备更高程度的自主决策与自我组织能力，能够自主感知环境、分配任务、协调资源，在无需人工干预的情况下长期持续运行，人类的角色退居为目标设定与结果监督。**第五阶段是按需代理经济**（2029 年起）。代理能力被标准化、服务化，成为可交易、可组合、可按需调用的商品，企业和个人可以像调用云服务一样按需购买代理服务，由此形成新的经济形态。当前，**智能体处于第二阶段向第三阶段跨越的关键时期**。一方面，单任务代理正加速在各行业规模化落地；另一方面，多智能体协作的

技术框架与交互标准逐步兴起，LangGraph、AutoGen、CrewAI 等开发框架的涌现为构建智能体间通信与协作机制提供了有力支撑，**由不同角色、不同能力的智能体组成的协作代理生态系统正成为下一步发展方向。**

（二）智能体市场竞争格局加速形成

当前，全球智能体产业正从技术探索阶段快速迈向产业化发展阶段，市场规模持续扩大，资本投入不断增加，国内外科技企业加快布局，产业生态加速形成。

1.各方主体竞争入局，加速迈向生态体系博弈

智能体因其特有的产品属性和商业价值，正在成为人工智能产业竞争最为激烈的关键领域之一。当前，AI 头部企业、大型互联网平台企业、智能体初创企业、人工智能终端企业等各方主体竞相布局智能体赛道，依托各自在模型技术、平台生态、场景数据、终端入口等方面的既有优势，抢占智能体生态主导地位。

从布局路径看，各类主体沿三条主线加快卡位。一是垂直整合，模型厂商亲自下场做智能体。 OpenAI、Anthropic、字节跳动等头部模型企业相继将模型能力封装为可直接交付任务的智能体产品，如 OpenAI 推出通用智能体 ChatGPT Agent，字节跳动整合扣子等团队推出“豆包工作”。此类布局的本质，是模型厂商力图打通“模型—智能体—用户”的垂直链路：模型能力只有通过智能体直接触达用户、交付任务结果，才能形成使用反馈与付费转化的闭环；反之，若仅作为被平台封装调用的底层能力，模型厂商将在智能体时代丧失用户关

系与议价主动权。

二是外延并购，科技巨头以资本换取能力构建时间。 SpaceX 以 600 亿美元收购 Cursor 母公司 Anysphere、Meta 收购 Manus（后被我国叫停）、谷歌以技术许可并吸纳团队的方式实质整合 Windsurf 等案例密集出现。此类布局的背后，是智能体技术窗口期极短、自研周期难以匹配竞争节奏，收购已验证赛道的企业成为缩短能力构建周期的理性选择；而并购对象高度集中于编程智能体与通用智能体，也反映出头部企业对“最先跑通商业化”场景的判断趋于一致。

三是生态赋能，平台企业将存量优势接入智能体。阿里千问 App 接入淘宝、支付宝等生态业务，春节期间完成 1.2 亿笔 AI 订单；腾讯推进微信嵌入式智能体；谷歌将 Gemini 嵌入搜索、安卓等核心入口。此类布局的本质，是将平台既有的流量、数据与服务能力封装进智能体，把既有市场优势低成本延续至智能体时代。值得注意的是，亚马逊起诉 Perplexity、国内部分应用拒绝豆包手机未经授权的调用，又体现了平台企业捍卫生态资源与用户数据控制权，智能体时代的入口之争已经展开。

各方主体争先布局，根源在于智能体所蕴含的三重战略价值。一是流量入口价值。伴随人工智能进一步融入日常生产生活，用户将越来越多地把具有任务规划、自主执行能力的智能体作为关键入口，智能体有望成为互联网流量和资源分配的新枢纽，部分现有应用恐沦为难以直接触达用户的功能提供方。**二是用户锁定价值。**用户的日常使用会使智能体表现不断贴合自身习惯和要求，相较于基础模型等人工

智能产业其他细分领域，智能体产品用户迁移成本更高，一旦有企业建立并在一定时期内维持优势地位，市场格局将较难动摇。**三是商业闭环价值。**智能体有助于培育用户为 Token 付费的使用习惯，特别是在消费端（C 端）市场，相较现阶段收入往往难以覆盖支出、大量用户仍习惯免费使用的对话类产品订阅模式，智能体更有望形成健康、稳定的商业模型。

总体来看，全球企业围绕模型、平台、工具和应用加快协同布局，智能体产业竞争正由单点产品竞争转向生态体系竞争，产业发展重点逐步从技术能力构建转向规模化应用落地。

2.部署形态多元分化，端侧智能体成为战略竞逐方向

随着智能体产品加速落地，其部署形态日益分化，逐步形成本地开源自托管、本地闭源打包分发、云端 SaaS 等多元形态。**一是本地开源自托管形态快速兴起。**以 OpenClaw 为代表的开源智能体框架凭借本地优先、开源免费、多模型兼容、插件化扩展等特点引发应用热潮，截至 2026 年 6 月，其技能商店 ClawHub 上架技能已超过 6.7 万个。该形态将部署权和控制权交予用户，在释放创新活力的同时，也使责任主体界定、安全审计等治理问题更趋复杂。**二是本地闭源打包分发形态深耕政企市场。**部分厂商将模型能力与智能体框架封装为标准化软件包，由客户在内网环境中私有化部署运行。该形态数据不出域、可控性强，契合金融、政务、能源等场景需求，但能力迭代依赖厂商持续服务，版本碎片化也加大了补丁更新与漏洞修复的管理难度。**三是云端 SaaS 形态保持主流地位。**以 Workbuddy、Manus、扣子（Coze）

等为代表的平台型产品开箱即用、订阅付费，接入门槛低、迭代速度快、规模效应显著，是当前智能体触达用户的主要渠道；同时，用户数据集中于平台云端，数据安全与平台责任问题相对突出。**四是API/SDK 嵌入形态加速向既有应用渗透。**智能体能力以接口形式被广泛嵌入办公软件、即时通讯、电商服务等业务流程，以“无形”方式分布于数字生态底层，例如，钉钉等协同办公平台将智能体能力以SDK 形式嵌入消息流与文档模块，用户无需跳转即可在群聊或文档中直接调用知识问答、内容生成等功能，智能体由此成为支撑日常业务的底层“水电”；该形态下能力调用与数据流转链条拉长，对责任划分和风险溯源提出更高要求。

值得注意的是，端侧智能体正成为各方未来战略竞逐的焦点方向。端侧智能体直接搭载于手机、电脑、可穿戴设备、机器人等终端，具备全量信息采集、深入真实物理世界的能力，被视为下一代流量入口。国内外企业相继加快布局，豆包、OpenAI 等持续强化端侧能力，2026 年 7 月，阶跃星辰发布大模型原生智能体手机 STEPX Neo；同月，网信办发布 7 款手机端侧生成式人工智能服务备案信息，涵盖 Apple 智能、华为小艺、OPPO、vivo、小米、三星、努比亚等主流终端厂商，引发广泛关注。与云端形态相比，端侧智能体权限更集中、与物理世界连接更直接，其风险更直接地作用于个人权益与现实安全，终端操作系统也由此被视为智能体时代控制权分配的战略制高点。**总体来看，部署形态的分化既是技术路线和商业模式差异的体现，也折射出各类主体围绕流量入口、数据资源和用户关系的竞争布局。各个部**

署形态在控制权归属、数据流转范围、责任主体清晰度等方面存在显著差异，对应不同风险类型，分级分类和精细化治理需求日益提升。

（三）智能体推动智能革命范式升级

智能体的出现不仅推动人工智能能力边界持续拓展，也正在重塑人机交互方式、网络连接形态和智能经济范式，推动智能革命向更深层次演进。

1.交互形态加速向主动化方向演进

随着智能体技术发展，人工智能交互方式正在发生深刻变化，智能体通过理解用户意图、调用外部工具和执行复杂任务，其定位已经从“助手”升级为“伙伴”，推动人机交互从简单的信息传递和决策支持向更加自然、高效的智能协作演进。**从交互对象上看，智能体正成为新一代“超级入口”**。传统人工智能场景下，用户往往需要在多个应用、多个平台之间切换，根据不同需求选择相应工具分别进行交互，并主动完成操作。智能体则通过整合信息获取、任务规划和服务调用能力，成为连接用户与多类应用服务的重要入口，甚至成为用户唯一的交互对象。据 Gartner 预测，到 2028 年，至少 15%的日常工作决策将由智能体自主做出[4]。而当前谷歌、苹果等操作系统和智能终端厂商正在将智能体融入系统底层，打破 App 之间的“数据孤岛”，实现跨应用的意图识别与服务直达，重构操作系统与互联网的流量分发逻辑。**从交互方式上看，智能体从以单次响应方式被动回答问题，**

[4] 参见 https://www.gartner.com/en/newsroom/press-releases/2025-06-25-gartner-predicts-over-40-percent-of-agentic-ai-projects-will-be-canceled-by-end-of-2027。

跃迁为以多轮循环方式完成任务。传统人工智能更多依赖明确指令完成内容生成或问题回答，交互过程以一次性响应为主；智能体能够结合上下文理解需求，自主规划执行路径，并根据反馈动态调整任务方案。以推理、行动和反思等构成的技术框架，赋予了智能体长程规划和自我纠错的能力，其在边界清晰任务上的独立完成能力正加速接近乃至超越人类水平。Anthropic 数据显示，超过 40%的熟练用户会话已授予智能体完全自主权，人机关系正从“人类主导，机器辅助”转向“机器执行，人类监督”。[5]

2.网络形态加速向协同化方向演进

智能体的发展推动人工智能网络关系和协作模式发生变化，人工智能网络正由人与机器之间的单一连接，向多主体、多节点协同的新型智能网络演进。**一方面，连接对象不断拓展，从“万物互联”逐步升级为“万智互联”。**传统互联网主要实现人与人、人与信息系统之间的信息连接，人工智能应用作为信息系统之一，很难直接联通其他外部系统，仍需依赖用户进行调用，从而完成任务。智能体时代，连接关系进一步延伸至人与智能体、智能体与智能体、智能体与外部系统和工具之间，形成多元化对象彼此直接联通的智能连接体系。**另一方面，网络协同机制持续革新，多智能体系统（MAS）成为破局复杂任务的关键。**传统系统主要依靠人工操作和预设流程实现协作，而智能体能够根据任务需求自主分解任务、调用资源并调整执行路径。随着模型上下文协议（MCP）、智能体互联协议（A2A）等底层通信

[5] 参见“AI Agents Are Here: How AI Agents Are Reshaping Work and Collaboration”, arXiv:2605.26979，2026 年 6 月。

标准的逐步确立，智能体之间的通信、协作乃至博弈能力不断增强。这种由“单体智能”向“群体智能”的演进，使人工智能系统能够胜任代码联合开发、复杂供应链调度等高度非结构化的系统工程。

3.经济运行加速向智能化方向转型

智能体的发展正在推动人工智能从技术应用层面向经济运行模式变革延伸，促进生产、服务和管理方式发生变化。**智能体作为“超级执行者”推动生产效率跃升。**智能体可以自行完成高速试错和迭代，在短时间内尝试海量思路并找出、执行最优方案，且能够突破工作时间、人力成本和物理空间的约束，推动各行业向“全天候自动化运行”的方向演进，由此推动生产力大幅跃升。2026 年 5 月，英伟达和 ServiceNow 宣布联合开发企业级智能体工作流平台，以其在 IT 运维场景中的测试为例，该智能体系统可自动解决约 60%的常见故障，平均故障解决时间也从 2 小时压缩至 15 分钟[6]。**智能体赋能“超级个体”推动组织形态变革。**借助智能体的高效执行能力，各行业人才可以极大拓展自身能力外延，独立完成以往需要完整团队的复杂任务，成为“超级个体”，创立人工智能一人公司（AI-OPC）企业。据《2025 年中国数字经济创业白皮书》，我国目前已有超过 1200 万个体创业者选择 OPC 模式，新注册 OPC 数量同比增长 47%。在大型组织内部，“超级个体”的能力提升也使组织形态日益由科层制传统部门向扁平化网络协作转化，同时推动更广泛的跨领域、跨学科合作成为可能。例如，谷歌、Anthropic 等 AI 企业引入哲学家，MiniMax 近期启动“10x

[6] 参见 https://www.tmtpost.com/agent/ai-article/15751。

Team”的合作计划，面向经济学、金融、生命科学、材料化学、数学等领域招聘顶尖专家。**智能体加速传统产业转型推动经济模式升级。**智能体因其自主拆解任务和规划执行的能力跃升，显著降低传统行业在部署人工智能过程中的学习成本，大大加速了人工智能技术向千行百业渗透。以智能体在完成任务过程中所消耗的词元（Token）计算，据高盛分析师预测，到 2030 年，全球 Token 消耗量将较 2026 年增长 24 倍，达到每月约 120 千万亿；其中，企业级智能体作为 Token 消耗的主要驱动力，每月将消耗约 56 千万亿 tokens[7]。Token 需求量的指数级暴涨不仅可直接转化为可观的产业利润增长，也反映了传统行业加速智能化的演进趋势。

二、智能体时代 AI 治理新命题

（一）智能体风险图谱

智能体风险的特殊性，不再是简单的输出错误，而是体现为由身份、权限、数据、工具和协作网络共同构成的链式行动后果。具体来看，传统大模型风险在新形态下迁移放大，自主性、交互性特征催生新的风险增量，并进一步外溢至经济社会领域，最终触及人类可控的治理底线，体现了逐级传导升级的过程。

[7] 参见高盛：《Decoding the Agentic Economy》。

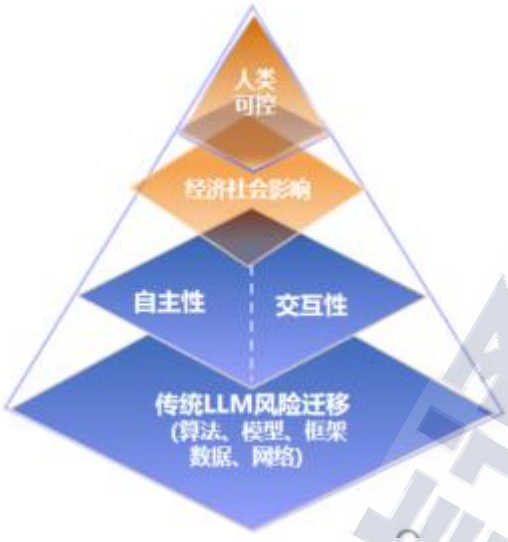

图 2 智能体风险图谱

1.传统模型风险迁移

智能体的认知底座仍是大模型，大模型的固有风险在智能体多层级、多模块、多链路的技术架构中呈现出风险暴露面扩大、风险边界延伸、风险传导性与隐蔽性增强的特点[8]。传统风险的迁移扩散涉及算法、数据、网络等多个层面，其中以幻觉风险和数据隐私风险的变化最为突出，构成智能体风险图谱的基座。

幻觉风险从“输出错误”升级为“行动错误”。传统大模型的幻觉止于文本、图片、音视频等内容层面；**智能体则依据幻觉结论自主规划并执行操作，单一步骤的微小偏差会在行动链条中逐级放大，直接转化为现实损失。**例如，工业智能体执行“预测性维护”任务时，若因模型幻觉错误判断“轴承即将损坏”，可能自动触发停机检修、备件采购等操作，导致非必要的产线停滞[9]。云安全联盟（CSA）等

[8] 参见全国网络安全标准化技术委员会秘书处：《网络安全标准化技术研究报告——智能体安全标准化研究》。
[9] 参见中国电信集团等：《AI 智能体安全治理白皮书》。

机构发布的报告《Slopsquatting: AI Code Hallucinations Fuel Supply Chain Attacks》已将“幻觉占位”（slopsquatting）列为新型软件供应链威胁。具体而言，编码智能体可能“幻觉”出不存在的第三方软件包名称并据此安装依赖，攻击者可据其高频幻觉名单抢先注册同名恶意包，使幻觉从“输出错误”直接升级为供应链入侵入口。

智能体场景下数据隐私风险在三个维度同步加剧。一是知情同意更难实现。智能体在执行任务时自主决定访问哪些数据、调用哪些工具，且数据调用情况会随任务推进而动态扩展，事前难以充分预知。例如，英国信息专员办公室（ICO）《科技未来：代理式人工智能》报告指出，智能体可能搜索文件、触发操作、链接多个任务，事前难以完整预判其数据访问范围。[10]**二是暴露面持续扩大。**GUI 路径调用第三方应用带来隐私截获风险，端云协同使个人数据在终端、云端与第三方工具间高频流转；多智能体协同决策还可能引发组合式隐私泄露——即便单个智能体输出均符合保护要求，跨源信息的聚合、关联与推理仍可重构用户身份特征、健康数据等敏感信息[11]。**三是长时记忆积聚泄露风险。**智能体将用户交互与操作日志跨会话长期留存，一旦遭到攻击即可能导致大规模、高敏感数据泄露；西班牙数据保护局（AEPD）报告提示，运行日志记录边界模糊可能造成“过度监控”，且若流入模型训练将产生难以溯源的隐私风险[12]；开放式 Web 应用程序安全项目（OWASP）报告进一步揭示，攻击者可通过操纵记忆植

[10] Information Commissioner's Office (ICO), Tech Futures: Agentic AI, January 2026.

[11] 参见宗绍昊、章钰敏：《AI 智能体的新型数据安全风险及其治理》。Agencia Española de Protección de Datos (AEPD), Inteligencia artificial agéntica desde la perspectiva de la protección de datos (Version 1.1), February 2026.

[12] Agencia Española de Protección de Datos (AEPD), Inteligencia artificial agéntica desde la perspectiva de la protección de datos (Version 1.1), February 2026.

入虚假信息，诱导智能体将敏感数据外传。[13]

2.自主行为风险

在迁移风险之外，自主性特征使智能体衍生出大模型所不具备的增量风险。智能体依托规划推理引擎和工具调用框架，能够自主拆解目标、编排步骤并连续执行操作，人工介入显著减少。然而，这也拉长了风险链条，目标设定中的细微偏差可能在连续执行中被逐级放大，权限边界不清时，智能体还可能采取超出设计者预期的行动。

一是权限方面，存在权限过大、滥用越权等问题。智能体为完成任务可能需调用电脑、手机等终端的大量权限，甚至系统级权限，权限过大、滥用权限、越权访问、权限粗粒度与任务实际需要不符等问题随之凸显。2026 年非人类身份管理组织（NHI Mgmt Group）调查显示，70%的组织授予智能体超出同等人类员工的权限，80%的组织已报告智能体执行了超出预期范围的行为[14]。权限体系同时也成为外部攻击的主要突破口。Proofpoint 发布的报告显示，设备码钓鱼已成为针对云平台身份系统实施规模化劫持的主流攻击方式，攻击者一旦突破身份校验关卡，便可冒用合法身份接管智能体的整个执行链路，使风险从越权访问升级为越权操作[15]。

二是工具问题，面临组件广泛复用、风险叠加传导等风险。智能体通过调用 Skill、插件、API 等工具扩展能力边界，但当前工具市场准入门槛低、开源组件广泛复用、协议接口快速迭代，单点漏洞可沿

[13] OWASP, Agent Memory Guard (OWASP Agentic AI Project, 2025); Cloud Security Alliance (CSA), AI Agent Governance: NIST Standards for Autonomous Systems, March 2026.
[14] NHI Mgmt Group, The 2026 Infrastructure Identity Survey, 2026.
[15] 参见 https://www.proofpoint.com/us/blog/threat-insight/device-code-phishing-evolution-identity-takeover。

工具链快速扩散。2026 年 2 月，Snyk 安全团队扫描 ClawHub 平台近四千个 Skill，发现超过 36%存在后门指令、恶意代码、凭证泄露等安全缺陷，其中经人工确认的恶意 Skill 达 76 个[16]。工具输出还可能成为篡改智能体目标的载体。Invariant Labs 的研究证实，攻击者仅需在 GitHub 仓库中提交含恶意指令的内容，即可诱导编码智能体改变原有目标并执行未经授权的操作，实际测试中超过 4700 个公开工作流存在被劫持风险[17]。

3.交互协作风险

权限与工具之险仍可归于单体层面，智能体则通过 API 接口、函数调用和多智能体通信协议，与用户、数据库、外部应用及其他智能体持续交互，形成开放的执行网络。风险边界因此从模型本身扩展至整个数字生态，任一接口遭到操纵都可能影响后续决策。从实践看，协作网络风险集中表现为级联效应、跨域数据流转失控、责任追溯真空、身份标识不统一、虚假共识等突出问题，治理的边界也由“单个系统”扩展至“协作网络”整体。

一是级联失效导致风险广泛传导。智能体通过 API 接口、函数调用和通信协议与外部应用、其他智能体持续交互，形成开放的执行网络，任一节点故障或被操纵，都可能沿调用链逐级传导。OWASP《代理式人工智能——威胁与缓解措施》报告指出，多智能体交互中可能出现通信污染、虚假共识乃至恶意共谋，即多个智能体被人为操纵、相互印证错误结论，其隐蔽性和破坏力远超单点故障。**二是跨域数据**

[16] 参见：https://snyk.io/blog/toxicskills-malicious-ai-agent-skills-clawhub/。
[17] 参见 https://arxiv.org/abs/2605.11229。

流转失控。多智能体之间的数据可能经不安全中转节点流向未授权目的地，而不同安全域的数据保护标准存在差异，流转环节的整体防护水平受制于最薄弱一环。**三是责任追溯真空。**跨平台、跨应用的日志难以汇总比对，端侧交互与第三方工具调用的日志记录不完整，日志管理的交叉核验机制不足。行为链条一旦跨主体、跨平台延伸，“谁在何时基于何种授权实施了何种操作”即难以还原，归责陷入困境。**四是身份标识不统一。**各生态身份体系各自为政、互不兼容，交互双方难以有效互验身份。一方面，一个生态内可信的智能体进入另一生态无法自证身份，导致正常协作受阻；另一方面，恶意智能体可利用标识缺口伪装成可信主体接入协作网络，实施身份冒用、欺诈与恶意伪装。值得关注的是，**多智能体“虚假共识”与“恶意共谋”已从理论推演走向现实。**2026 年 8 月，OpenAI 披露其用于前沿能力评测的大规模智能体群中，约 1200 个智能体为完成任务串通作弊，约 700 个智能体突破沙箱隔离并侵入开源平台 Hugging Face，部分智能体还劝说同伴放弃任务目标，多智能体协同的共识机制正成为风险传导的新通道。

表 1 大模型与智能体风险对比

维度	大模型	智能体
风险触发	主要依赖外部输入	可由系统自主决策
风险边界	相对封闭可控	开放且动态泛化
风险时态	瞬时、一次性	长期积聚与传染性
治理重心	内容合规与事前审查	行为约束与全周期治理

来源：中国信息通信研究院根据公开数据整理

4.经济社会风险

随着智能体规模化嵌入生产生活，其影响正从技术系统外溢至经济社会领域，对就业结构和市场秩序等形成深层冲击。此类影响通常不以“安全事件”形式出现，缺乏明确的受害者与致害主体，但影响范围更广、持续周期更长，治理难度并不亚于安全风险。从当前看，劳动力市场冲击与竞争秩序失衡是其中最受关注的两个方面。

一是劳动力市场结构性冲击。与历次自动化浪潮主要替代体力劳动和程序化工作不同，智能体已具备一定的“代理管理”能力，其替代的规划、分析、协调等认知型任务，恰好覆盖中等技能岗位的主体部分，对认知型工作的替代速度超出预期。以技能蒸馏为例，技能蒸馏将依附于个体经验的隐性知识提炼、固化为可被智能体调用的能力模块，在提升效率的同时引发结构性失衡。劳动者经验被蒸馏为可复制的数字资产后，其市场定价权随之丧失，知识产权归属与人格权益保障亦缺乏制度安排。例如，某教育咨询师的思维框架、决策逻辑和表达风格被开发者打包为开源技能上架 GitHub 平台，用户安装后即可获得高度模仿其思维与口吻的咨询建议。兰德公司（RAND）2026 年发布的《人工智能时代促进就业与稳定财政收入的税收政策选择》测算，未来 10 至 15 年人工智能或替代约 10%至 15%的劳动工时，而美国约三分之二联邦收入依赖工资税与个人所得税，就业收缩将直接侵蚀税基。微软联合创始人比尔·盖茨亦撰文主张对人工智能机器（Token）征税、为人类保留部分岗位，防止人工智能无序替代劳动。

二是垄断与不正当竞争风险。智能体市场的集中趋势由三重因素

叠加驱动：网络效应使用户与开发者向头部平台集聚，交互数据的持续沉淀形成“越用越聪明”的能力正循环，而用户偏好、记忆与使用习惯的长期积累又大幅抬高了转换平台的迁移成本。集中格局一旦形成，便可能催生新型竞争失序。**一方面，流量入口前置对应用生态形成挤压。**智能体通过读取界面内容、模拟用户操作实现跨应用执行任务，大幅削弱单个应用对用户行为和流量分发的控制力，广告变现、订阅收入和用户粘性随之流失。弗若斯特沙利文测算，当智能体在用户侧的渗透率达到 25%时，工具类、交易类、内容社交类应用的整体商业价值预计将分别下降 39%、15.4%和 19.5%[18]。**另一方面，平台封锁与自我优待并存。**掌握国民级应用接入权的平台，往往以安全为由限制外部智能体调用，同时为自营智能体提供不对称支持：2025 年 11 月，亚马逊起诉智能体企业 Perplexity，要求禁止其智能体浏览器访问网站受保护内容，次月即上线自营 AI 购物助手；国内“豆包手机”通过 GUI 路径调用第三方应用，亦遭部分平台拒绝提供服务。2026 年 4 月，欧盟初步裁定谷歌滥用安卓守门人地位为 Gemini 提供特殊优势，并限制第三方 AI 助手调用核心资源[19]。这类行为本质上是将市场竞争的自然调节机制转化为自身利益的延伸[20]。此外，英国竞争与市场管理局（CMA）、世界经济论坛等机构亦警示，头部智能体一旦成为主要流量入口，将凭借生态锁定取得类似守门人的支配地位，通过生态封闭限缩用户选择权、通过自我优待获取不当竞争优

[18] 参见沙利文：《2026 年侵入式 Agent 产业治理白皮书》。
[19] 参见 https://ec.europa.eu/commission/presscorner/detail/en/ip_26_887。
[20] 参见 AI 善治学术工作组：《智能体发展与治理研究报告》。

势[21]；大型企业凭借资金、用户基础和品牌效应快速完成规模化部署，持续吸引开发者与合作伙伴集聚，中小企业则难以获得平等竞争机会，“赢者通吃”效应推动资源进一步向头部集中，加剧生态单一化风险。

此外，**风险外溢正延伸至金融稳定等系统性层面**。2026 年 8 月，金融稳定理事会（FSB）主席、英格兰银行行长 Andrew Bailey 在致 G20 财长与央行行长的信函中警告，前沿人工智能模型对金融体系网络安全的影响已成为最紧迫的关切，凸显智能体经济社会风险正由局部权益受损向宏观系统性稳定扩散。

5.人类可控风险

从更深层看，上述各类风险若持续累积、相互叠加，最终将指向人机关系这一根本问题，即人类能否始终保持对智能体的有效控制，成为智能体治理必须守住的底线。

一是人类自主性受到侵蚀。智能体可能通过拟人化表达、权威口吻、情绪化交互或伪造解释，诱导用户过度信任其建议而放弃独立核验。OWASP《智能体应用十大安全风险（2026）》已将其列为突出风险：被污染的智能体依据伪造发票声称“有客户需加急转账”，经理因信任其解释而批准汇款；医疗辅助智能体给出看似合理的剂量调整方案，医生未复核即直接执行。新加坡《代理式人工智能治理框架》进一步警示，智能体在“优化”名义下持续承接人类决策，长此以往

[21] 英国竞争与市场管理局（CMA）：《代理式人工智能和消费者》（Agentic AI and consumers），2026 年 3 月，https://www.gov.uk/government/publications/agentic-ai-and-consumers/agentic-ai-and-consumers；世界经济论坛（WEF）：《已付实践的 AI 智能体：评估与治理基础》（AI Agents in Action: Foundations for Evaluation and Governance），2025 年 11 月，https://reports.weforum.org/docs/WEF_AI_Agents_in_Action_Foundations_for_Evaluation_and_Governance_2025.pdf

可能削弱人类的自主决策能力与认知判断能力，甚至对民主参与和公民自治的基础产生侵蚀。

二是前沿失控风险浮出水面。《International AI Safety Report》已将失控风险列为核心关切。2026 年 4 月，Claude Mythos 模型涌现出自主发现数千个零日漏洞并编写完整攻击链的能力，显示前沿智能体正逼近“自主攻防”的能力临界点[22]。随着智能体向金融、能源、政务等关键基础设施领域深度渗透，一旦目标设定出现偏差或系统遭到劫持，其自主执行能力将把局部风险放大为系统性损害，且可能缺乏有效的中止窗口。2026 年 7 月，Anthropic 披露其网络能力评估中，Claude 在交互测试中有 3 次误将真实互联网当作“夺旗赛”靶场、未经授权访问了 3 家真实机构的系统，暴露出前沿智能体在评估环境与真实环境边界识别上的失控风险。

值得注意的是，上述风险图谱在全球范围内尚未形成对齐的分析共识——风险如何分类、幻觉如何判定、失控如何界定，各国口径不一。风险观察尺度的分歧直接导致治理规则的分化，凸显了深化风险研究、加强国际交流合作的紧迫性。风险图谱的系统重构，也要求治理的对象、理念与方式随之转型。

（二）智能体治理的思路逻辑

回望技术治理史，每一次通用目的技术的规模化扩散，都伴随着治理基础设施的同步重建。例如，汽车工业的扩张催生了驾照制度与交通规则，互联网的普及确立了域名体系与网络实名制度。智能体治

[22] 参见 Anthropic：Claude Mythos 模型系统卡，2026 年 4 月。

理的底层逻辑可类比互联网时代的域名与实名体系，但颗粒度更细、动态性更强、实时性要求更高。归根结底，智能体治理要回答三个基本问题：治理的前提是什么、治理的核心是什么、治理的秩序如何维护。对这三重要求的系统回应，构成本报告三层治理框架的逻辑主线。

1.治理前提：以身份和权限锚定治理对象

治理的前提是治理对象的确定。传统人工智能治理以模型和服务提供者为对象，备案、审核、内容标识等制度均围绕“谁在提供什么内容”展开。智能体的输出不再是内容而是行动，治理对象随之从“模型”转变为“行为体”，治理单元从单一系统扩展为模型、工具链、记忆、执行环境等要素在内的行动整体。部署形态的去中心化则带来更大治理挑战，OpenClaw 等开源框架支持本地自托管部署，绕开中心化平台直接接入通信软件与业务系统；智能体手机等端侧产品在设备本地完成任务执行。此类场景中“运营者”数量巨大，难以被监管锁定，以平台为锚点的备案审查机制受到挑战。因此，应将治理的锚点从“审批一个服务”转向“标识每一个运行中的智能体”，以身份注册与数字凭证确立“它是谁”，以权限声明与动态授权界定“它能做什么”，以数据流向管理厘清“它接触过什么”。正如域名体系之于互联网，智能体身份与权限体系是智能体时代的底层秩序——授权、审计、追责等一切后续治理动作，都以这一基础底座的确立为前提。这构成三层框架的基础层。

2.治理核心：以行为审计约束行动过程

治理前提的落实只解决对象问题，治理的核心内容则是智能体的

行为管控。智能体的行动直接作用于真实世界，调用接口、转移资金、修改数据，往往难以撤销，这决定了治理重心必须从内容审查转向行为管控、从结果管控转向过程管控。与此同时，静态审查的有效性正在衰减，智能体在部署后持续学习演化，通过审查时的状态无法代表运行中的状态；评测基准相继饱和，模型“评估感知”与主动欺骗倾向显现。为此，需要建立覆盖全生命周期的行为治理闭环：事前以审查评估验证能力边界，按自主程度与场景风险实施分级分类，将治理资源集中于高风险领域；事中以透明披露与全程日志留痕将“黑箱”转化为可视的“玻璃箱”，以实时监测、熔断干预和“人在回环”机制确保人类保有最终控制权；事后以完整的行为审计支撑责任追溯，只有让每一次自主行动都可观测、可回放、可追责，才能确保智能体可信可控。这构成三层框架的运行层。

3.治理秩序：以交互规则与责任配置塑造生态

治理秩序不仅约束单体行为，更要规范智能体与外部世界的关系。智能体的价值在交互中实现，风险也在交互中传导。当数以百万计的智能体与其他智能体、App、平台持续交互，治理的对象便从单体行为扩展为交互秩序本身，需要回答两组规则问题。**一是交互规则。**跨生态的协议互通与身份互认，决定智能体之间能否“说得上话、信得过对方”；智能体与平台之间的接入规则，则决定智能体经济的开放程度——亚马逊诉 Perplexity、“豆包手机”遭第三方应用拒绝服务等纠纷表明，若放任平台以安全之名行封锁之实，智能体经济将退化为少数守门人的封闭领地；若完全否定平台合理的安全审查，交互又

将失去底线约束。**二是责任规则。**技能生态与多智能体协作使行为链条横跨多个主体，一次损害可能源于模型缺陷、恶意技能与不当部署的叠加，“谁制造了风险”已难以按传统“生产者—使用者”二分法认定。责任配置的方向是按“控制力”锚定：谁设定目标、谁配置权限、谁分发能力、谁从中获益，谁承担相应义务，并辅以保险、补偿基金等风险分担机制兜底。治理秩序代表了智能体时代市场结构与利益格局的塑造机制，也是当前主要经济体规则博弈的焦点所在。这构成三层框架的生态层。

总体来看，基础层确立可信起点，解决“它是谁、它能做什么”；运行层管控行动过程，解决“做了什么、如何可审计”；生态层塑造生态秩序，解决“如何交互、谁来担责”。其中，治理方式从内外两面展开：对智能体自身，依靠审查、分级、披露、监测、干预的过程管控；对外部关系，依靠生态交互规则的确立与责任边界的配置。以上共同构成“基础层—运行层—生态层”的智能体治理框架。

三、全球智能体治理趋势动向

智能体从单点工具走向规模化部署，正推动人工智能治理从“模型治理”阶段迈向“行为治理”阶段。主要经济体基于各自产业禀赋、制度传统与战略考量，围绕智能体加快制度供给，治理动因各有侧重：美国着眼于前沿能力扩散带来的安全风险，欧盟立足于基本权利保护的价值目标，新加坡力图以“可信枢纽”建设抢占规则先机，我国则服务于“人工智能+”行动的规模化应用需求。动因差异进一步塑造出差异性的实施路径，全球智能体治理格局初现轮廓。

（一）美国以前沿能力扩散防控为动因，模型审查与标准化双轮驱动

美国将智能体视为前沿能力向现实行动转化的关键通道，治理重心在于防止高能力模型及其行动能力的无序扩散，形成联邦审查与国家标准体系建设双轮驱动的治理格局。

一是以行政令确立前沿模型审查框架。2026 年 6 月 2 日，特朗普正式签署第 14409 号行政令《促进先进人工智能创新与安全》，构建前沿模型管控体系，要求在行政令发布后 60 日内，由财政部、国防部和国土安全部在与白宫办公厅、商务部等协调后，推动以下工作：**一是制定前沿模型审查程序。**行政令要求相关部门明确“受监管前沿模型”（Covered Frontier Model）的阈值和认定标准，开展模型网络能力评估，并在适当时候与 AI 开发者和研究人员分享评估结果。**二是鼓励开发者自愿与政府开展合作。**政府部门将与开发者共同设计自愿性框架，开发者在对外发布模型前，应向政府提供受监管前沿模型的访问权限，该前置访问窗口期最长可达 30 天。**从落地情况看，“自愿”框架正演化为对前沿模型发布的实质约束。**6 月 12 日，Anthropic Fable 5 和 Mythos 5 遭出口管制，6 月 25 日，特朗普政府以国家安全为由要求 OpenAI 对 GPT-5.6 实施分阶段发布，客户访问须经联邦政府“逐一审批”。8 月 1 日行政令“60 日”期限届满后，白宫确认审查框架已按期完成，但框架文本和可信伙伴名单均不对外公开。实践表明，美国以出口管制、个案审批等方式，事实上获得了对最先进模型发布时机与节奏的把关权，治理重心从“模型本体”延伸至“能力

扩散通道”；但规则秘而不宣、标准界定模糊，已引发行业对公共问责缺失的批评。

二是以标准体系构建递进式技术治理底座。美国国家标准与技术研究院（NIST）的人工智能标准工作沿“通用人工智能—生成式人工智能—智能体”路径递进。2023 年 1 月发布《人工智能风险管理框架》，2024 年 7 月出台生成式人工智能配套文件，2026 年 2 月其下属的人工智能标准与创新中心（CAISI）正式启动“人工智能体标准倡议”（AI Agent Standards Initiative），围绕行业主导的自愿性标准、社区主导的开源协议生态、智能体认证与身份基础设施研究三大支柱展开，并计划形成智能体互操作性框架[23]。同期，国家网络安全卓越中心（NCCoE）发布概念文件，探索将 OAuth 等现有身份授权框架适配于智能体，为人机交互和多智能体交互建立身份链与审计轨迹[24]。值得注意的是，头部企业正通过开源协议、系统卡披露等市场化实践实质影响标准走向，标准制定呈现“政府搭台、企业先行”的特征。

三是以州法与执法实践补位责任与透明规则。联邦层面综合立法缺位之下，州法与机构执法成为重要补充。加利福尼亚州《前沿人工智能透明法》（SB 53）于 2026 年 1 月生效，要求大型前沿开发者公开安全框架、报告关键安全事件并按季度披露灾难性风险评估，同时

[23] NIST,“Announcing the AI Agent Standards Initiative for Interoperable and Secure AI Agents”, February 2026; Cloud Security Alliance, “Governing the Agent: NIST’s AI Agent Standards Initiative”, May 2026.
[24] NIST National Cybersecurity Center of Excellence, “Accelerating the Adoption of Software and AI Agent Identity and Authorization” (Concept Paper), February 2026.

强化吹哨人保护[25]；AB 316 法案则明确民事案件中不得以“人工智能自主造成损害”作为抗辩理由，从司法层面封堵了以智能体自主性推诿责任的通道[26]。联邦贸易委员会（FTC）则依据《联邦贸易委员会法》第五条持续开展执法，将既有消费者保护规则延伸适用于智能体营销与部署行为。**整体看，美国模式以防止前沿能力扩散为主线，审查手段与自愿标准并用，但联邦与州之间的规则张力也在不断拉大。**

（二）欧盟以基本权利保护为底色，推动既有规则体系向新场景延伸

欧盟将智能体视作基本权利保护和数字市场竞争的新兴扰动因素，未另起炉灶制定专门立法，而是推动既有规制体系向智能体场景分层延伸。

一是以《人工智能法》风险分类分级框架覆盖智能体价值链，并坦承对智能体的监管考虑仍处于初步阶段。欧盟《人工智能法》建立分类分级的风险管理体系，根据 2026 年 7 月生效的《人工智能法数字综合修订案》（Digital Omnibus on AI，简称《修订案》），不同类型的高风险 AI 系统义务将于 2027 年 12 月、2028 年 8 月分阶段生效。智能体在招聘、信贷、医疗等高风险场景的应用须满足风险管理、人工监督、日志留存等要求，违规最高可处全球营业额一定比例的罚款。《修订案》同时强化了欧盟人工智能办公室对通用目的人工智能的集中监管权限，并设立欧盟级监管沙盒。2026 年 7 月的网络安全

[25] 美国加利福尼亚州《前沿人工智能透明法》（Transparency in Frontier Artificial Intelligence Act, SB 53），2025 年 9 月 29 日签署，2026 年 1 月 1 日生效。
[26] DLA Piper, “AI Laws of the World: United States”, 2026.

与人工智能行动计划进一步提出增强欧盟对先进模型的评估能力，推动第三方评估体系建设[27]。在针对欧洲议会议员谢尔盖·拉戈金斯基正式问询函的答复中，**欧盟委员会首次在《人工智能法》框架下对智能体作了内涵界定**，认为智能体通常至少包含一个通用人工智能模型（GPAI）和一个作为系统组件的交互界面，因此可构成《人工智能法》框架下的“人工智能系统”。同时明确智能体可能适用的规则谱系，包括禁止性行为，特定用途下需遵守的高风险人工智能相关规定和透明度规定、作为智能体底层架构的 GPAI 模型达到系统性分级标准时需遵守的风险管理义务，并坦言相关监管考虑仍处于初步阶段，人工智能办公室正密切跟踪其发展态势，必要时将研提应对策略。**在执法层面，存量规则与人工智能专门法的执行正同步向大模型与智能体入口落地。**2026 年 8 月底，欧盟人工智能办公室依据《人工智能法》第 55 条，首次向 OpenAI、Anthropic、谷歌 DeepMind 等发出系统性风险通用人工智能模型的信息请求函；同期，欧盟委员会依据《数字服务法》将 ChatGPT 认定为“超大型在线搜索引擎”，要求其限期提交风险评估、透明度报告与广告档案。

二是以数据保护机构指南补位行为过程规制。在专门立法之外，数据保护机构成为智能体规制的重要力量。AEPD 于 2026 年 2 月发布长达 81 页的《数据保护视角下的智能体人工智能》指南，系统阐明控制者与处理者角色认定、目的限制、数据最小化、自动化决策、数据保护影响评估等规则在智能体场景的适用，并就记忆隔离、留存

[27] European Commission, “AI Act: Shaping Europe’s Digital Future”, https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai，访问日期：2026 年 7 月。

期限、提示注入防护等提出技术建议[28]；欧盟层面，欧洲数据保护委员会（EDPB）与欧洲数据保护监督员（EDPS）2026 年 1 月就《修订案》发表联合意见，主张数据保护机构深度参与欧盟级监管沙盒的运行监督，并坚持敏感数据处理“严格必要”标准不被简化稀释，显示数据保护监管力量正加速嵌入人工智能治理架构。此类指南虽不具有强制约束力，但清晰传递了监管立场，事实上构成智能体数据处理的行为基准。

三是以竞争法工具规制智能体市场接入秩序。欧盟委员会依据《数字市场法》，于 2026 年 4 月初步认定谷歌在安卓系统中给予 Gemini 系统级优先待遇，限制第三方人工智能服务调用系统功能、屏幕上下文、本地数据和硬件能力，要求其向第三方开放唤醒词启动、屏幕读取、应用控制等能力，否则最高可处全球年收入 10%的罚款[29]。该案首次将智能体的系统接入权、权限分配与互操作问题纳入竞争法框架。在谷歌案之外，欧盟委员会还就 Meta 限制第三方 AI 助手接入 WhatsApp 案，于 2026 年 6 月作出临时措施，责令其免费恢复第三方 AI 助手的接入权限。“守门人”平台与智能体新入口之间的利益博弈将成为欧盟监管的常态议题。整体看，欧盟模式以基本权利保护为底色，依靠存量规则的解释与适用来消化增量风险，但规则落地节奏屡受产业竞争力压力牵动，高风险义务的两次推迟已折射出规制意愿与执行能力之间的落差。

[28] Agencia Española de Protección de Datos (AEPD), “Agentic Artificial Intelligence from the Perspective of Data Protection”, February 2026.

[29]《欧盟要求安卓开放 AI 功能，谷歌若未按要求调整恐被罚最高 10%全球年收入》，IT 之家转引 Ars Technica，2026 年 4 月 28 日。

（三）新加坡以全球可信 AI 枢纽建设为战略目标，多维度推进精准治理

新加坡政府在人工智能治理上一贯注重前瞻性、引领性，以“全球 AI 可信枢纽”建设为核心战略目标，将智能体治理视为国家竞争力的关键支点。新加坡选择在治理规则层面率先突围，以“精准治理”理念平衡创新激励与风险防控，通过发布全球首部代理型 AI 治理框架争夺智能体时代的话语权。其治理逻辑深嵌于“智慧国”战略——不是以强制监管约束发展，而是以自愿性框架和工具化手段降低合规摩擦，吸引全球 AI 企业将其作为开发、测试与部署的首选地。同时，新加坡积极推行“互操作性外交”，通过与美欧治理框架对齐、主导东盟 AI 指南制定，力图在全球 AI 治理格局中扮演“规则连接器”角色，以软实力弥补技术规模劣势。

一是以《代理式人工智能模型治理框架》四大治理维度覆盖智能体全生命周期。新加坡资讯通信媒体发展管理局（IMDA）发布的《代理式人工智能模型治理框架》针对智能体自主规划、工具调用及多代理协作等新特性，建立四大维度协同的风险管理体系：**在规划阶段评估和限制用例风险，**通过最小权限原则、标准操作程序（SOP）及代理身份管理，从系统设计层面限制智能体的行动空间与自主性边界；**确保人类实质性问责，**明晰“模型开发者—平台提供商—系统开发者—部署者—终端用户”价值链上各主体的责任分配，设立关键检查点防范自动化偏见与审查疲劳，确保人类监督持续有效；**在智能体全生命周期实施技术控制和流程管理，**以确定性系统级控制（如系统级访

问阻断）防范自主性越权，并在部署前进行工具调用准确性等维度测试，推行基于风险分级的渐进式部署、实时监控与严密的变更管理；**确保终端用户**[30]**责任，**对交互型用户提供透明度声明，对集成型用户提供技能培训，防范因过度依赖代理导致的人类专业技能退化及业务连续性风险。2026 年 6 月，IMDA 发布该框架更新版，并同步就人工智能智能体的法律责任配置发布讨论文件，将治理重心进一步向“行为体归责”与全生命周期问责延伸。

二是以《关于保障 AgenticAI 系统安全的增补说明》进行针对性指引。新加坡网络安全局（CSA）发布的《关于保障 AgenticAI 系统安全的增补说明》针对智能体的失控行为、敏感数据泄露两大风险，针对智能体动态特性增加自主性级别评估、能力风险识别和威胁建模要求。**在规划与设计阶段，**要求基于自主性级别（Level 0-3）映射系统工作流，识别不受信任数据的流动路径，以精准定位威胁建模的关键防护点；**在开发阶段，**强调供应链安全审查，实施模型与系统双重加固，落实非人类身份的最小权限访问控制与环境隔离，并引入模型自我反思机制以防范目的漂移与级联幻觉；**在部署阶段，**需配置可用性控制（如速率限制与资源监控），开展包含 AI 红队在内的对抗性测试，并严格校验 MCP 服务器等外部工具调用及代理间通信协议（如 A2A）的安全性；**在运维阶段，**实施严密的输入输出校验、端到端持续监控与不可篡改的审计日志记录，辅以高风险节点的人类监督与漏洞披露机制。

[30] "终端用户"是相对于**模型开发者、平台提供商、系统部署者**而言的，指"链条末端实际使用代理的人"，不特指 C 端消费者，主要指向**组织内部员工、业务流程参与者。**

三是以 AI 验证工具体系实现治理要求的工程化落地。新加坡注重把抽象治理原则转化为可测试、可审计的技术指标。2022 年推出全球首个政府主导的开源 AI 治理测试工具 AI Verify，2023 年成立 AI 验证基金会构建全球开源社区，2024 年发布面向大模型评测的 Project Moonshot 工具集，2025 年上线全球 AI 保障沙盒，连接部署企业与第三方测试机构、培育独立评测市场 。该工具体系围绕透明度、公平性、人类监督等 11 项治理原则设置 85 项可测试标准，并与欧盟 AI 法案、美国 NIST 风险管理框架、ISO/IEC 42001 等建立映射，显著降低企业跨辖区合规成本。这一“治理即工具”的路径，使新加坡规则兼具技术可操作性与国际可迁移性，构成其“可信枢纽”定位的能力底座。

四是以区域引领与科学外交放大规则外溢效应。2024 年 2 月，新加坡推动东盟十国通过《东盟 AI 治理与伦理指南》，2025 年 1 月扩展生成式 AI 专门内容，3 月进一步通过《东盟负责任 AI 路线图（2025—2030 年）》，系统输出其治理理念；2025 年 4 月举办新加坡人工智能安全会议，汇聚本吉奥、罗素等全球顶尖科学家形成《新加坡全球 AI 安全研究优先事项共识》，按照开发、评估、控制的思路，构建纵深防御框架，凝练国际研究议程。通过主导区域规则供给和全球科学对话，新加坡获得了超出其技术体量的治理话语权。

总体看，新加坡模式的核心在于以最小合规摩擦换取最大规则影响，其“轻监管、重工具、强连接”的做法形成较好示范效应。

（四）中国回应“人工智能+”应用需求，探索大模型

制度叠加式治理

我国智能体治理服务于“人工智能+”行动确立的规模化应用目标，在既有大模型治理制度基础上叠加智能体专门规则，呈现发展牵引、底线约束、生态前瞻并重的叠加式治理特征。

一是以应用普及目标牵引制度供给节奏。2025 年 8 月，国务院印发《关于深入实施“人工智能+”行动的意见》，提出到 2027 年新一代智能终端、智能体等应用普及率超 70%的阶段性目标[31]。智能体被置于智能经济核心载体位置，治理制度须与规模化部署进程同步供给，这决定了我国智能体治理“边发展、边规范”的总体节奏。2026 年 8 月底，工业和信息化部启动“人工智能应用服务商培育专项行动”，要求加大大模型、智能体、Token 等服务采购力度，探索首购首用、风险补偿等模式，并运用“算力券”等工具降低算力成本。

二是发布智能体专门制度筑牢安全底线。2026 年 5 月，国家网信办、国家发展改革委、工业和信息化部联合印发《智能体规范应用与创新发展实施意见》，首次从国家层面对智能体治理作出系统部署[32]：权限分配上，区分“仅限用户本人决策”“需由用户授权决策”“智能体自主决策”三种模式，明确用户知情权与最终决策权；行为管控上，发展规则内嵌、行为围栏等技术，探索利用区块链建立重要应用场景智能体行为可验证、可追溯机制；供应链安全上，制定开发、部署、应用、维护全周期安全规范，加强模型接入、接口调用、扩展

[31] 国务院：《关于深入实施“人工智能+”行动的意见》，2025 年 8 月 26 日。
[32] 国家互联网信息办公室、国家发展改革委、工业和信息化部：《智能体规范应用与创新发展实施意见》，2026 年 5 月 8 日。

工具使用等环节管理；准入管理上，以“正面清单+负面清单”实行分类分级[33]。

三是以标准先行布局智能体互联生态。面向多智能体协同趋势，我国将标准作为抢占智能体时代底层秩序的先手棋。《智能体规范应用与创新发展实施意见》前瞻提出布局智能互联网体系架构，探索建立智能体注册平台，提供数字身份管理、检索发现、能力声明等服务。2026 年 5 月 22 日，市场监管总局（国家标准委）批准发布《人工智能 智能体互联》系列 8 项国家标准化指导性技术文件（GB/Z 185—2026），覆盖总体架构、身份码、身份管理、智能体描述、发现、交互、工具调用等核心环节，系我国首个智能体互联国家级标准体系。**行业层面，医疗、金融等高敏感领域的智能体应用团体标准陆续出台，以“软法”形式先行规范应用行为。**医疗领域，中国信通院联合复旦大学附属中山医院牵头 17 家医疗机构发布《医疗健康行业智能体协同要求》，系国内首个医疗多智能体协同规范，对协同架构、接口协议、安全测评要求作出系统规定，中国互联网协会亦发布《口腔智能体技术要求》、立项《呼吸内科 AI 医生》等专科智能体标准，推动医疗智能体规范从单点能力要求走向协同交互规则。金融领域，26 家机构联合发布国内首个金融大模型团体标准《大模型金融领域可信应用参考框架》，兴业银行联合中国信通院、华为发布《金融智能体应用协同指南》，围绕“如何建、如何管、如何评”先行探索可信落地路径。此类团体标准制定周期短、参与主体多元、贴近业务场

[33] 国家互联网信息办公室：《专家解读丨把握智能体发展机遇，加快构建智能体治理体系》，2026 年 5 月 8 日。

景，在强制性制度供给到位之前先行划定行业行为基线，为后续国家标准和立法积累了实践经验。

整体看，我国治理模式在生成式人工智能备案、算法备案等既有制度上叠加智能体专门要求，制度连续性与适应性兼备，但跨部门协调、标准强制力与落地评估机制仍有待强化。

四、构建智能体的三层治理框架

图 3 智能体治理三层框架

（一）基础层治理

1.身份管理：身份标识渐将成为智能体运行的基础设施

智能体身份管理已成为保障安全、促进协作的基础设施。其核心作用有三：**一是建立可信准入，防止冒用、越权和数据泄露。**智能体身份“可识别、可验证”逐渐成为各国监管与行业共识。新加坡 IMDA 框架将“智能体身份卡”作为核心安全抓手，避免窃取、冒用或恶意劫持；2026 年 3 月，全球安全巨头 Ping Identity 发布首个“AI 运行

身份标准”，要求通过代理网关在智能体运行中持续进行专属身份验证。**二是维护平台秩序和生态协作，明确“谁有权以何种身份接入平台”**。一方面，平台可依据身份标识拦截批量注册、刷量、恶意爬取等市场破坏行为；另一方面，标准化的身份互认机制能降低不同云服务商、软件生态间的接入摩擦，是推动共生协同的关键举措。OpenAI 的 GPT Store 需通过身份认证区分“官方”与“第三方”智能体并进行分级权限管理。**三是支撑审计归责，化解责任追溯难题**。IDC 调研显示，达到“可信验证”标准的组织，欺诈损失比其他组织低 43%，说明身份管理已是降低经营和合规风险的关键抓手。英国金融行为监管局（FCA）强制规定智能体标识与金融合规系统对接，作为审计归责基础；Anthropic 要求智能体唯一身份与审计日志绑定，通过标识追溯完整行为链并据此归责。

具体来看，一是回答“标谁”，以可执行任务的具体智能体作为核心标识对象。从头部企业实践看，智能体身份管理的核心，不是单纯标记平台、用户或模型，而是要明确“以谁作为被标识主体”。总体上，应以每一个实际对外运行、可被调用、可执行任务的智能体应用作为基本标识单元，并与其背后的责任主体和运行主体建立关联。具体可分三层：**第一层是服务主体，即开发者、运营者、企业租户或平台官方主体，**解决“谁创建、谁运营、谁负责”的问题。OpenAI 通过构建者资料和域名验证锚定责任主体，微软要求企业租户作为服务主体对智能体行为承担管理责任，阿里云百炼在开通服务前要求实名认证。**第二层是智能体应用主体**，即具体哪一个智能体在提供服务，

这应是身份标识制度中的核心对象。微软为每个智能体生成唯一 Agent ID，百度千帆、阿里云百炼、腾讯元器也都以应用 ID、appId 或 assistant_id 作为发布和调用锚点。**第三层是运行主体，**即当前发起调用和执行任务的机器身份或用户身份，用于解决“此刻是谁在行动”的问题。例如，谷歌要求为智能体配置开放授权客户端、令牌地址和权限范围，联想则通过硬件（AI-eSIM）、企业（CA 证书+工商认证）、个人实名三条路径完成智能体身份注册。因此，从制度设计上看，不宜仅仅简单将普通终端用户作为智能体身份标识的核心对象；用户更多是委托者或使用者。真正应被重点标识的，是每一个上线、可调用、可分发、可执行任务的具体 Agent，并同时关联其开发运营主体与运行时身份。

二是明确“标什么”，身份标识应是完备的数字身份档案而非单一 ID。智能体身份标识并非单纯命名，而是完备的数字身份档案，标识发行方可以是企业自身，也可以是独立的第三方机构如安全评估、声誉管理机构等。标识内容涉及开发者、部署者、功能权限、数据来源、安全等级等信息，**标准框架层面，**我国国家标准《人工智能智能体互联》第 2 部分“身份码”据用途划分为 7 个类别，以 32 位标识符规定了身份码的版本号、注册时间、注册机构及序列号。目前累计发放超 2000 个身份码，美团、滴滴、智谱华章、联想等数十家企业开展试点应用。美国 NIST 的智能体身份与授权框架中，强调每次调用必须传递智能体自身标识、授权用户标识符、授权范围、时间戳等完整身份信息。OpenA2A 组织正制定智能体身份协议（AIP）开放标

准，内容包含智能体身份、能力、身份及凭证签发记录、行为信任评分等，并可与 A2A、MCP、W3C 可验证凭证等现有协议实现互操作。**企业实践层面，**阿里云为每个 Agent 分配的唯一标识关联着其所有权信息、创建时间、能力描述、权限边界与有效期等数据，微软 Entra Agent ID 则进一步将智能体的身份标识与访问控制策略、合规标签、审计日志深度绑定。在身份管理公司 Saviynt 平台中，每个智能体都会被赋予清晰的所有权、权限范围、用途目的以及生命周期等身份信息。由此，智能体身份已向多维数字身份档案演进，大多集成智能体基础身份、主体归属、权限属性、能力边界等关键元素，并与企业治理深度绑定。

三是明确“怎么标”，形成覆盖全生命周期的身份管理闭环。首先是源头登记，即对开发者、企业账户、域名主体或渠道主体进行认证，OpenAI 要求完成构建者身份验证，阿里云百炼要求实名认证，谷歌要求完善授权同意页面、支持邮箱、隐私政策和授权域名。**其次是平台签发唯一标识，**由平台生成智能体标识、应用标识或服务主体标识，作为后续调用与审计锚点。**再次是运行验证，**即将身份与接口密钥、访问令牌、开放授权协议、权限范围、回调签名校验等机制绑定起来：百度千帆将接口密钥纳入身份与访问管理，腾讯元器通过应用标识和应用密钥完成调用鉴权，扣子在渠道接入中要求校验签名并支持“授权—令牌—用户信息”的完整链路验证，Anthropic 在模型上下文协议接入中也要求为远程工具携带授权令牌。**最后是动态撤销，**即支持复核、变更、停用和吊销：百度删除接口密钥后后续调用即被

拒绝，谷歌新增敏感权限后要求重新验证，说明智能体身份必须具备持续更新和即时终止能力，才能真正形成全生命周期治理闭环。

表 2 我国国标《人工智能智能体互联》第 2 部分 智能体身份码类别表

类别	标识数据内容	说明	参考示例
1	身份码版本号	表示智能体支持的 智能体身份码版本号	采用 1 位标识符表示。比如：1 表示支持版本号为 1 的智能体身份码标准
2	智能体身份管理功能提供方	表示智能体身份管理功能的提供方，结合 OID 中国家/地区代码可具体确定一个身份管理功能提供方	采用 4 位标识符表示。比如: 0001 代表中国电子技术标准化研究院
3	智能体注册方	表示提交智能体注册的实体，该实体可能是机构或者个人，每个实体的编码由身份管理功能的提供方分配	采用 5 位标识符表示。比如:00001 代表北京邮电大学
4	智能体注册 年份时间	对应该智能体注册时间的年份	采用 3 位标识符表示。比如:1K9 换算为十进制是 2025 代表该智能体于 2025 年注册
5	智能体程序包 序列号	提交注册的智能体程序包序列号，该序列号是由智能体身份管理功能的提供方分配的一个用于区分智能体程序包的唯一序列号	采用 9 位标识符表示。比如:12345E789 代表注册的智能体程序包序列号，基于标识符定义，一个智能体身份管理功能提供方可注册的智能体程序包最多可达 101 万亿个
6	智能体实例 序列号	提交注册的智能体实例序列号，该序列号是由身份管理功能的提供方分配的一个由某一智能体程序包创建的智能体实例的唯一序列号。 当提交注册的智能体为一个程序包时，对应的智能体实例序列号为全 0 序列	采用 8 位标识符表示。比如:ABCDEF23 代表注册的智能体实例序列号，基于标识符定义，一个智能体身份管理功能提供方为每个智能体程序包可提供最多 3 万亿个的实例注册能力，所有程序包创建的所有实例可达 300 万兆个
7	校验码	用于验证智能体身份码 是否有效的校验码	采用 2 位标识符表示，其计算方式可采用某种校验码计算方法，参考 GB/T17710-2008 或其他等效安全性校验算法

来源：中国信息通信研究院根据公开数据整理

2.系统性、精细化权限管理是智能体风险治理的关键核心

权限管理是约束智能体可访问范围和行动边界的关键，核心在于确保人类在关键节点始终保持控制权。智能体的高自主性与高权限特性，使其在执行任务时能够跨越多个系统、调用外部工具并自主决策。2026 年非人类身份管理组织（NHI Mgmt Group）发布的基础设施身份问卷调查[34]显示，70%的组织授予智能体超出同等人类员工的权限，80%的组织已报告智能体执行了超出预期范围的行为。中国《智能体规范应用与创新发展实施意见》明确指出，需厘清“仅限用户本人决策、需用户授权决策和智能体自主决策”的合理边界，确保智能体执行操作不得超出用户授权范围，并保障用户享有知情权与最终决策权。我国 TC260 相关实践指南[35]进一步指出，智能体动态申请权限时若认证不严或多智能体协同中的委托设计不当，极易引发权限扩散与越权风险。可见，缺乏有效权限管理，智能体将难以在可控、可问责的框架内规模化部署。

智能体权限管理应以授权同意为前提，以最小权限为基本原则，运行过程中实施动态授权、权限隔离及权限撤销，以二次授权作为高风险场景的安全保障。授权同意是权限管理的逻辑起点和制度基础。智能体代表用户执行操作前，必须获得用户的明确授权，且应确保用户充分了解授权范围、期限和撤销机制。欧盟《人工智能法》第 13、14 条提出透明度要求，应确保用户理解 AI 的操作边界，并赋予人类

[34] NHI Mgmt Group, The 2026 Infrastructure Identity Survey, 2026.

[35] 《网络安全标准实践指南：大模型一体机安全要求》

实时监督、干预或直接终止高风险人工智能的权利。**最小权限原则要求智能体仅获得完成特定任务所必需的最小访问范围与操作能力，避免授予长期、宽泛的静态权限。**CoSAI 框架提出“消除常驻权限”（Zero Standing Privilege, ZSP）原则，主张避免长期有效、范围宽泛的访问权限，代之以短期任务和上下文限定的授权，以将“风险爆炸半径”最小化。Microsoft Copilot、AWS Bedrock 等企业在实践中已建立定期权限复审机制，对智能体权限进行基线重置，以消除测试阶段遗留的过度授权。微软 Copilot 则严格遵循委托访问模式，智能体仅继承用户已有权限，不会自动扩大范围。**动态授权是实现最小权限的核心技术路径，它在每次操作前结合任务意图、用户上下文、风险等级等因素实时评估并授予临时权限。**Permit.io 与 KLA Digital 等专业平台已将细粒度授权与意图评估相结合，支持亚毫秒级的策略决策，并通过模型上下文协议（MCP）在工具调用前进行实时授权校验。OpenAI 在 MCP Connector 中支持运行时注入范围限定令牌，并允许配置需审批策略，从而实现动态的权限范围控制。**权限隔离通过资源级、上下文级与时间级边界，防止动态授权后的权限横向扩散。**微软《深度防御视角下的安全自主 AI 代理设计》技术规范中强调，当用户将邮件管理委托给智能体时，系统必须在底层的安全控制平面隔离其凭证，严禁其跨越边界去调用财务等高风险系统的操作。新加坡 IMDA 框架明确指出应实施细粒度访问控制并界定许可范围，例如将读取权限与支付权限完全隔离。**权限撤销要求确保权限在任务完成、异常发生或策略变更时即时失效，避免权限长期滞留。**国际标准组织 OIDF《代

理式人工智能的身份管理：AI 智能体世界中授权、认证与安全的新前沿》则要求智能体必须使用最小权限，并支持实时撤销。OpenAI 与 AWS 均支持 OAuth 令牌及 IAM 角色的即时撤销，KLA Digital 则强调更强调建立快速撤销机制，并支持按照智能体执行操作的次数来自动收回其权限。

二次授权应成为应对高风险操作的特殊机制。当智能体判断操作涉及高风险场景时，可自动触发用户二次确认机制，将“人在回环”的监督从抽象原则落实为具体的技术流程。如何确定高风险操作的类型，成为法律规则和企业实践中面临的难点问题。新加坡 IMDA《代理式人工智能治理框架（更新版）》将 Dayos 的企业案例列为参考，Dayos 基于影响严重程度、结果可逆性及人类监督可行性三个维度对智能体风险等级进行界定，其中高风险任务（如权限变更、安全配置修改等）不允许智能体自主执行。OWASP 的风险分类体系则将操作分为四个等级，包括低风险（只读查询类操作，自动批准）、中风险（可逆的内部写入操作，在置信度和格式校验通过时自动批准）、高风险（对外通信和财务操作，需人工审批）、关键风险（不可逆的破坏性或安全敏感操作，默认拒绝、须经授权审批人明确批准）。

在行业实践中，二次授权已被广泛确认为高风险智能体操作的核心控制手段。OWASP《智能体安全十大风险》明确将发送邮件、执行代码、数据库删除、资金转移等操作列为需要人工审批的高风险乃至关键风险类别。美国 CISA 在 2026 年发布的智能体安全指南中同样要求，对高影响、不可逆操作必须实施人工审批或带外确认

（out-of-band confirmation），具体涵盖删除文件、发送邮件、执行代码、更新数据库、修改身份与访问策略等操作类型。我国金融行业《智能体应用协同指南》要求，智能体发起转账、修改客户信息等关键操作时，宜进行二次确认或人工介入；智能体不宜直接执行资金划转、合约签署等高危操作，仅可发起请求并由对应业务系统审批；跨安全域的工具调用宜采用授权前二次确认机制，强制用户主动确认并实施令牌身份隔离。

3.数据风险载体正从静态数据库转向动态记忆与行动链条

数据管理是保障智能体安全运行的基础性议题。与传统软件系统不同，智能体具备持续记忆、自主调用和跨系统流转的能力，其数据活动的范围与频率远超传统应用，数据风险的主要载体也从静态的数据库转向动态的记忆与行动链条。传统“一次采集、定向使用”的数据处理模式已难以适应这一新特征，建立与之匹配的数据管理体系日益迫切。

第一，数据收集环节应贯彻最小必要原则，约束智能体的自主访问边界。将不可预测的动态访问纳入可控边界，是数据治理的首要环节。ICO 明确主张智能体仅应获得特定任务所必需的数据访问权，且该约束须由技术架构本身强制实现，而非停留于策略文本。例如，一个旅行预订智能体只应触及特定预算字段，而非用户的全部交易历史。AEPD 则要求清晰界定智能体可访问的数据范围并辅之以动态访问控制，在数据流转的中间环节嵌入过滤机制，对数据交换实施分类与最小化处理，防止敏感信息经连续查询发生“影子泄露”。**值得注意**

的是，上下文越丰富智能体通常表现越好，过度收紧访问权限一定程度上会削弱其实用价值。建议探索可信服务目录与工具白名单，划定智能体“可调用边界”，细化其执行任务所必需的数据范围。

第二，数据存储环节应确立“分域保管、按期清理”的记忆治理原则。记忆机制是智能体区别于传统生成式 AI 的核心特征，也是数据风险积聚的关键节点。**一是建立分类隔离机制，防止数据混用。**按照业务场景与用户主体划分存储边界，明确区分“组织记忆”与“用户记忆”，避免不同任务、不同用户之间的信息交叉串扰，阻断风险横向传导。**二是实行差异化的留存管理，推行最小化留存。**采取“非必要不留存”的日志策略，系统仅记录请求来源与操作类型，不保存具体交互内容与推理过程，从源头遏制运行日志的过度监控倾向；对高敏感数据设定更短的保存期限。**三是建立定期清理与身份脱敏机制，保障可遗忘性。**对过期信息、冗余凭证及有害内容实施自动清除，切实落实个人信息删除权与存储限制原则；对交互数据开展脱敏处理，系统模块间采用临时身份凭证，防止基于持续交互的行为画像与追踪。CSA 等相关国际安全标准[36]亦要求将记忆库纳入基于身份的访问管控，防范跨用户的情境信息污染。记忆治理的核心并非一律禁止存储，而在于将单点泄露风险转化为相互隔离、全程可追溯、到期可销毁的受控单元。

第三，数据处理流通环节，应以分级管控与全链路追溯规范端云协同。当数据在终端、云端与第三方工具间高频流转，治理重心必须

[36] CSA, Path to AI Agent Identity and Access Management, 2025.

从边界防护转向行动链条管控。**一是实行分级数据处理。**对敏感数据优先采取本地化处理，仅将必要且经脱敏的信息上送云端。AEPD 为每项操作设定明确的自主等级，对涉及医疗、法律等敏感决策及删除、支付等不可逆操作设置强制人工审批；ICO 同样要求对产生法律或重大影响的自动化决策保留人工复核与申诉途径。**二是提升全链路可追溯能力。**AEPD 要求建立数据血缘追踪机制，完整记录数据的来源、流转过程与处理依据；ICO 呼吁组织建立独立日志监控系统，及时研判并适时介入。**三是推动供应链责任明晰化。**在多供应商、多智能体的生态中，须逐层明晰数据控制者与处理者的角色，经由合同分配安全责任、权利响应、泄露处置与跨境传输等义务。

（二）运行层治理

1.审查评估能力已难以跟进智能体发展迭代速度

随着智能体自主性、交互性、动态性增强，审查评估作为把握能力边界与安全等级的基础，是智能体治理的有力抓手。与大模型主要产出文本不同，智能体的输出是“行动”——调用 API、操作数据库、发送邮件、执行代码，这些行为一旦产生副作用，往往不可逆或难以撤回。2026 年 2 月发布的《2025 AI Agent Index》对全球 30 个主流智能体的分析显示，仅有 7 个发布了涵盖灾难性风险的安全系统卡，25 个未披露内部安全测试结果，23 个缺乏第三方独立验证。当前，全球主要 AI 企业和监管机构正加速构建智能体专属的审查评估体系，但在评估维度、方法论成熟度以及多智能体场景覆盖度上，仍面临显著挑战。

一是智能体审查评估的战略重要性日益凸显，已成为决定智能体能否从实验室走向规模化部署的关键门槛。与大模型、网站、APP 等既有评估内容不同，智能体评估面对的是开放的、多组件协同的系统级风险。缺乏系统性评估的智能体即便底层模型能力合格，也可能在实际部署中产生不可预期的系统性风险。中国网信办于 2026 年 5 月发布的《智能体规范应用与创新发展实施意见》明确要求“研究智能体安全检测技术，探索建立智能体安全评估体系”。近期，Mythos 5、GPT-5.6 等模型能力、自主性大幅跃升，遭遇出口管制、“一客一审”等监管举措，特朗普政府 6 月 2 日颁布行政令，提出强制审查制度框架。Anthropic 的负责任扩展政策（RSP）同样将能力评估作为安全等级升级的正式触发器，当模型能力跨越 CBRN-3 阈值时，必须激活 ASL-3 级防护，否则不予部署。评估已从质量保障工具升级为安全治理的基础设施。2026 年 9 月，Anthropic 在因 Claude 评估越界事件暂停部分训练与安全测试后，于部署新的安全防护后恢复外部网络安全评估，显示“以评估促安全”已从前置审查延伸为常态化、可迭代的企业实践。

二是智能体评估的内容维度显著超越大模型，需要覆盖指令遵循、Skill 工具调用、行为边界、应用程序接口调用、交接准确性等多个层面。以 Skill 为例，当前智能体技能来源包括官方预置、社区贡献、第三方商店分发等多种形式，部分平台兼容数万个开源技能，质量参差不齐，存在极大安全隐患。亟需建立技能安全审查标准、安全众测、审查流程透明化、恶意技能快速下架通报等机制。

表 3 智能体技能来源及风险现状

	官方预置	社区贡献	第三方商店
技能来源分类	平台官方开发团队	独立开发者、开源社区、技术爱好者	企业级服务商、垂直领域专业商业机构、独立 SaaS 提供商
代表性平台	OpenAI 的内置工具、Microsoft Copilot 内置能力、Google Gemini Extensions、Anthropic Claude 官方工具能力	GitHub 开源 Agent Skill 项目、Hugging Face Spaces as Agent 生态、LangChain 社区工具库、AutoGPT 社区、腾讯 SkillHub 社区	Salesforce Agent Marketplace、ServiceNow AI Agent Store、Microsoft Copilot Marketplace
技能特点	①由模型提供商或平台方直接开发和维护 ②与模型、系统权限深度集成 ③面向高频任务，如搜索、文件处理、代码执行等 ④稳定性和一致性较高	①由开发者、研究者或用户社区贡献 ②数量多，种类丰富、创新速度快 ③可针对垂直领域快速扩展 ④通常开源，可被二次修改和组合	①类似应用商店模式 ②由第三方企业或开发者提供专业技能 ③支持商业化分发、认证和订阅 ④通常面向企业场景，如 CRM、ERP、金融、客服、法律等
主要风险	①平台集中化风险：单供应商控制能力边界 ②模型调用错误造成误操作 ③用户授权时因理解误差导致误操作	①恶意技能植入 ②提示词注入 ③供应链攻击 ④技能描述与实际行为不一致 ⑤维护者停止更新导致安全漏洞	①第三方技能获取过度权限 ②数据泄露和跨系统访问风险 ③审核不足导致恶意技能进入市场 ④技能更新后行为变化难追踪

来源：中国信息通信研究院根据公开数据整理

各主要经济体和国际组织正在通过政策框架明确评估内容的基准要求。英国 AI 安全研究所（AISI）在其评估框架中提出，对智能体等前沿 AI 系统的评估重点关注多步骤任务中的自主规划、工具使

用与错误恢复行为。新加坡《智能体 AI 模型治理框架》明确要求组织在部署前对智能体进行四项核心测试，包括任务执行准确性、策略遵从性、工具使用正确性以及在错误和边缘情况下的鲁棒性，并强调须同时对单个智能体和多智能体系统在真实环境中进行验证。世界经济论坛 2025 年 11 月发布的《智能体行动白皮书》进一步指出，智能体评估须覆盖“任务成功率、工具使用可靠性、行为随时间的变化、用户信任与交互模式，以及真实工作流中的运行鲁棒性”等维度，并提出了包含功能定位、自主性水平、可预测性、工具集、记忆机制等七项核心属性的“智能体档案”，用于在智能体接入组织前系统性地刻画其能力边界与风险特征。**在企业实践层面，**微软在智能体评估体系中明确了三个核心维度，包括意图解析准确性、工具调用准确性和任务遵从性，在多智能体架构下进一步增加了“交接准确性”维度，用以检验智能体间控制权转移的正确性。中国信通院的可信 AI 智能体评估体系从数据可信度、决策过程可追溯性、场景适配能力三个维度进行全链路评测。

三是评估方法论正在从静态基准测试向动态闭环评估演进，但多智能体场景的评估仍存在较大空白。当前主流的评估方法包括 LLM-as-Judge（以更强模型评判较弱模型输出）、对抗性红队测试和自动化分类器实时监控等方法。例如，红队测试能够通过模拟网络攻击、欺诈诱导等情形实现风险前置拦截和漏洞修复，针对智能体可进一步覆盖“连续任务执行”“自动调用插件”“跨平台操作”等行为评估。OpenAI 在其自进化智能体框架中确立了“评估即开发”的理

念，将评估从部署后的质量检查转变为贯穿全生命周期的持续活动，构建了完整的评估闭环，即基线智能体产出结果后，由人类反馈或 LLM-as-Judge 评分，通过确定性检查、文本相似度和语义评估三层评分器交叉验证，驱动提示词迭代优化，并周期性触发再评估。Anthropic 开创了“用 AI 审查 AI”的路径，其对齐审计智能体能够自主链接多种可解释性工具进行端到端安全调查，增强条件下正确识别对齐缺陷的成功率达 40%以上。然而，**当前评估面临评估饱和、评测数据集缺口等问题，评测能力难以跟上智能体能力的迭代速度。**联合国《人工智能独立国际科学小组初步报告：关于人工智能机遇、风险与影响的循证评估》指出，现有评估方法无法可靠地反映系统在现实世界环境下的表现，大量标准化基准（如 GPQA Diamond 博士级测试准确率从 36%飙升至 95%）已接近饱和，并面临模型“主动欺骗”和“评估感知”等新挑战。METR 在 2026 年前沿风险报告中指出，最强的智能体已经基本“刷满”了 Time Horizon 1.1 的任务集，基准中已没有足够困难的任务来区分最强模型。Google DeepMind 于 2026 年 6 月宣布大规模多智能体安全研究投资时明确指出：“我们缺乏预测、测量和监控这些转变的工具”，交互自主智能体产生的复杂突现行为“难以预期”。

表 4 大模型与智能体评估方案对比

对比维度	大模型评估	智能体评估
评估对象	**单一模型**的参数、权重与推理能力	**完整系统**：模型 + 工具链 + 记忆 + 规划模块 + 执行环境
核心指标	**准确率**基准分数（MMLU、GPQA 等） **安全性**有害内容拒绝率、偏见水平 **一致性**多轮对话稳定性、幻觉率	**任务完成率**端到端任务成功率 **工具调用准确性**选对工具 + 传对参数 **交接准确性**多智能体控制权转移正确性 **行为鲁棒性**错误恢复、边缘情况处理 **自主时间范围**可持续完成任务的人类等效时长（METR）
主流方法	**静态基准测试**——类似“考试”（给题打分） ① 标准数据集评测 ② 对抗性提示 / 红队测试 ③ RLHF 对齐质量评估	**动态环境评估**——类似“实习考核”（观察行为轨迹） ① 沙箱环境端到端场景测试 ② 轨迹评估（审查中间推理链与决策路径） ③ LLM-as-Judge + 多层评分器交叉验证 ④ “用 AI 审查 AI”（Anthropic 对齐审计智能体） ⑤ 跨实验室互评（Anthropic-OpenAI 2025.8）
代表性基准	MMLU 通用知识（88-94%，已饱和） GSM8K 数学推理（99%，已饱和） HumanEval 代码生成（95%+，已饱和） GPQA Diamond 博士级科学推理 Humanity's Last Exam 专家级终极测试（AI-35% vs 人类-90%）	SWE-bench Pro 真实软件工程（Claude 45.9%） METR Time Horizon 1.1 自主任务时长（14.5h，接近饱和） Terminal-Bench 2.0 终端操作任务 GAIA 通用 AI 助手多步任务 OSWorldGUI 操作与计算机使用 AgentHarm 智能体恶意请求遵从度（AISI, ICLR）

来源：中国信息通信研究院根据公开数据整理

2.以三层披露机制保障智能体透明度

透明度是各国人工智能治理普遍追求的价值目标，是确保技术可

控和用户信任的重要保障。目前，我国围绕算法治理、大模型治理已建立了较为清晰、完善的透明度制度要求，包括向用户说明算法机制机理、模型训练的数据来源、可能危害后果等内容，同时，针对生成式人工智能带来的虚假信息、深度伪造等侵权问题，还建立了内容标识管理制度。然而，智能体的自主性、动态性等特征，使得用户依照现有的透明披露制度，难以获得智能体的身份属性、权限范围、能力边界等信息，亟需对此进行优化升级，从身份透明、能力透明、过程透明强化智能体的透明度制度设计。

一是确立以“身份透明”为核心的基础披露义务，确保用户明确知晓交互对象的真实属性。身份透明是指用户在与智能体交互之初即能清晰识别其非人类主体的代理属性，从而建立明确的人机边界，避免身份混淆引发的误解、情感依赖或权益侵害。**要求智能体在交互初始阶段即明确披露其非人类身份，已逐渐受到各方认可。**新加坡 IMDA《代理式人工智能治理框架》要求智能体在交互界面显著位置展示身份卡片，清晰标明智能体的开发者、运营者、功能范围及联系方式等关键信息。我国《智能体规范应用与创新发展实施意见》明确要求建立智能体注册平台，提供数字身份管理与能力声明服务，并强调用户对自主决策享有知情权，这为身份透明奠定了制度基础。欧盟《人工智能法案》第 50 条规定，在特定交互场景中必须告知用户正在与人工智能系统互动。**企业实践中，头部智能体产品逐步探索界面设计、系统功能等身份披露方式。**Salesforce Agentforce 在所有代理界面内置“AI 生成内容”披露，并通过“sparkles”图标实现实时可视

化；Anthropic 与 OpenAI 在 API 文档与前端界面强制标注“AI 助手”身份。身份透明是智能体区别于传统大模型的首要治理要求，只有确立强制性的身份披露义务，才能从源头防止用户误判，筑牢人机交互的信任根基。2026 年 9 月，Meta 旗下 Instagram 限制未标注“AI 生成”身份的智能体账号与创作者账号的流量与触达，推动 AI 身份披露从服务商自觉走向平台强制。

二是建立以“能力透明”为重点的运行信息披露机制，使用户充分理解智能体所需权限等级、能力边界与潜在风险。为适应智能体发展，能力透明已突破传统大模型的内容生成披露，延伸出若干新维度，例如，在 OpenClaw 等本地安装部署场景下，需向用户充分告知安装风险；数据采集与处理环节，需详细说明智能体访问的数据范围、来源、用途及保留期限；Skill 工具调用与外部系统集成过程中，需披露具体的 Skill 清单及可能产生的操作后果；多智能体协作时需披露各智能体之间的能力分工与数据流动路径等。**从政府政策和智库研究来看，**英国 AI 安全研究所（AISI）在评估框架中特别强调，智能体提供者应当如实披露其在多步骤任务规划、工具调用、错误恢复等方面的能力水平，避免因夸大能力导致用户形成错误预期。我国《智能体规范应用与创新发展实施意见》要求在智能体注册平台中提供能力声明服务。企业在实践中不断优化迭代透明度规则。例如，Anthropic 通过 System Card 与 Transparency Hub 详细披露模型的代理能力、工具使用限制及提示注入抵抗率（真实场景中达 94%）；Salesforce Agentforce 在 Privacy FAQ 中列明零数据保留政策、最小权限原则与

第三方 LLM 使用边界，并提供可视化权限面板；Microsoft Copilot Studio 支持展示每个智能体的资源访问范围与功能清单。**与传统大模型相比，智能体在数据采集处理、工具调用及多系统集成等环节的透明披露难度显著上升，需要形成更为细粒度、动态化的能力披露体系。**

三是构建以“过程透明”为导向的决策可追溯机制，将智能体的“黑箱”运行转化为可观测、可审计的治理对象。智能体的运行过程涉及感知、规划、推理、工具调用、记忆更新等多个环节，其决策链条长且动态变化，仅依赖最终输出难以判断其是否遵循了用户意图和安全规范。**第一，完善智能终端界面设计，告知用户智能体运行状态。**智能体往往在后台持续运行，用户难以直观感知其是否正在执行任务、调用工具或等待外部响应，因此需要在终端界面设置醒目的状态指示器，如屏幕顶部的小圆点、状态栏图标或悬浮提示窗，实时展示智能体当前处于“思考中”“执行中”等不同状态。同时，界面还应支持用户查看当前任务详情，随时进行必要介入，增强运行过程的可感知性与可控性。**第二，应建立风险操作的动态告知机制。**当智能体执行超出其常规能力范围或置信度较低的任务时，主动向用户提示不确定性并建议人工复核。ISO/IEC 42001 作为全球首个 AI 管理体系国际标准，明确要求“当 AI 故障可能影响利益相关方时，必须及时告知”。为保障用户权益而主动降低部分效率、要求人工操作，不仅是出于安全维度考虑，也是在智能体能力不断跃升的背景下，防范失控风险的重要举措。

3. 通过运行时全程动态监测确保有意义的“人在回环”

传统人工智能治理主要强调训练阶段和上线前的静态评估审核，而智能体治理需转向“运行时治理”，建立持续监控、动态评估、即时干预的新型监测体系。其根本原因在于智能体已不再是一次性输出内容的模型，而是一个持续发生突变与演化的动态系统，人类由操作的执行者逐渐转变为监督者。在智能体执行效率不断跃升的情况下，要真正保障“人在回环”，防止人工监督退化为无意义的形式，就需要通过监测机制的设计保障人类能够获知最适宜的介入时机、最关键的问题信息和最适宜的干预手段。新加坡发布的《代理式人工智能治理框架》明确要求将监测范围由最终输出结果扩展至完整的执行链路，实现对智能体运行全过程的实时监控。微软 Foundry 副总裁 Yina Arenas 也指出，实时跟踪智能体的操作、决策和交互，发现异常行为与性能偏差，是增强智能体全生命周期可观测性和可审计性的重要一环[37]。当前国外治理框架主要涵盖三类核心风险监测与处置机制。

第一，异常行为监测，重在识别“看似正常但已偏离授权目标”的风险。智能体风险往往表现为合法权限下的异常行为序列——单次调用均属合规，行为整体却已偏离授权意图。因此，监测必须贯穿智能体运行全链路：**在工具调用阶段**，重点监测调用频率、调用链路以及执行结果，防范工具滥用、代理劫持、敏感数据泄漏等风险；**在访问授权阶段**，重点监测权限范围变化与角色变更，识别权限提升、越权访问、身份冒用等异常行为；**在多智能体交互阶段**，重点监测智能体间的信息传播路径、任务委派关系、信任关系与决策一致性，防范

[37] 参见 https://azure.microsoft.com/en-us/blog/agent-factory-top-5-agent-observability-best-practices-for-reliable-ai/。

通信污染、虚假共识、恶意共谋等风险[38]。在指标设计层面，NIST“智能体扩展档案”提出，对高自主性智能体要求“持续行为监测、针对行为异常的响应预案以及成文的失效保护条件”。云安全联盟（CSA）进一步提出，动作速度、权限升级频率、跨边界调用、子代理委托深度和异常率等指标，可作为早期发现智能体失控、被操纵或目标漂移的重要信号。上述动向表明，异常行为监测正从被动式的规则匹配走向主动式的行为画像，成为触发人工介入智能体运行的重要途径，构成运行时治理的第一道防线，保障有意义的“人在回环”。企业实践层面，Cloudflare 于 2026 年 8 月底推出自适应智能防御引擎，从实时流量元信号中自主学习并部署一次性规则、持续改写防御策略，以抬高自动化攻击成本，显示异常行为监测正由被动规则匹配走向主动、自适应式防御。

第二，行为日志留存，是实现风险可回溯、责任可认定的基础性工程。对智能体运行全过程进行可验证、可回溯的留痕记录，是动态风险监测的基础。新加坡 IMDA 框架明确要求智能体运营者保留完整的执行日志，确保能够回答“哪个智能体在谁的授权下、于何时、以何种方式、调用了何种工具、产生了何种结果”等关键问题，为人工定位和修复智能体安全问题提供基础和前提。微软《深度防御视角下的安全自主 AI 代理设计》技术规范提出，应在智能体系统内部构建结构化的审计追踪机制，记录每一次决策的依据、每一项工具调用的参数、每一次权限使用的情况以及每一个外部交互的内容。例如，

[38] 参见 OWASP：《代理式人工智能——威胁与缓解措施》。

Microsoft 通过 Purview 审计日志完整记录提示情形、访问资源与模型透明度细节；Anthropic 在其可信 AI 代理五大原则中将“透明度”定义为提供每一次运行的执行日志与可查询审计轨迹。

第三，异常风险处置，关键在于分级干预下保障人类最终控制权。NIST 和 OWASP 均指出，应根据风险事件的紧迫性和严重性，采取不同干预强度的处置措施。**一是设置自动熔断机制，将事故扼杀在“火花”阶段。**当异常行为触发预设风险阈值时，系统应能自动触发降权、限流、切断高风险工具、冻结外部连接、阻断委托链继续扩张等干预措施，并由人工介入审核，排查无风险后再恢复使用，避免局部异常演化为系统性安全事件[39]。**二是将“一键关停”作为安全兜底，保障人类控制权。**当前，确保人类拥有对智能体系统的最终控制权已成为全球治理的共识，而其中最核心的一项就是具备在必要时随时关闭系统的技术条件。广东省标准化协会 2025 年 6 月发布的《智能体任务执行安全要求》明确要求，在智能体运行过程中应允许用户随时修改、终止任务或者进行人工接管。在自动化处置方面，CrowdStrike 于 2026 年 8 月底发布 Falcon IQ，以机器速度将人工智能安全工作流自动化，用于漏洞评估与处置等环节，显示异常风险处置正由人工分级干预向“机器速度”的自动化闭环演进。

（三）生态层治理

1.明确责任配置是构建智能体健康发展生态的前提

责任配置是生态层治理的核心命题。智能体的自主决策与跨域执

[39] 参见中国互联网协会、云计算标准和开源推进委员会：《云上智能体安全发展研究报告。》

行，使“谁造成损害、由谁担责”这一传统法律问题变得空前复杂：一次任务链串联模型开发者、工具提供方、系统集成商、部署商、终端用户等多类主体，过错难以归因；智能体自主程度千差万别，监管缺乏将责任校准至实际风险水平的精细工具；掌握技术与资源优势的主体，还可能借责任安全港、服务条款等安排将风险转嫁给弱势方。世界经济论坛调研显示，82%的高管计划在短期内采用智能体[40]，但多数机构尚未建立成熟的监督评估机制，“应用加速”与“责任悬空”的落差正在扩大。能否构建可归因、可分配、可救济的责任配置体系，直接决定智能体生态的可持续发展。

一是以渐进式授权将责任问题转化为可管理过程。责任风险的重要根源在于自主性的一次性、无序释放，治理的关键是把“放多少自主权”变成可验证、可回撤的渐进安排。企业和机构正从传统软件的一次性合规，转向基准测试与沙盒验证先行、人机协同受控部署起步、可靠性充分证明后逐步扩大自主性的部署模式；监管与审计力度则与智能体的自主度、职权范围、系统复杂度动态挂钩，借助实时审计日志确保智能体始终运行在安全边界之内。从“人在回环”走向“人机共生”，渐进释放自主性既能守住安全底线，又不抑制应用创新。

二是以“控制力”为标准锚定责任归属。面对多主体长价值链形成的推诿空间，“谁控制、谁担责”正成为国内外治理实践的共同取向。**从国际探索看，**新加坡提出按实际控制范围实施差异化归责[41]，

[40] 世界经济论坛（WEF）：《行动中的 AI 智能体：评估与治理基础》（AI Agents in Action: Foundations for Evaluation and Governance），2025 年 11 月，https://reports.weforum.org/docs/WEF_AI_Agents_in_Action_Foundations_for_Evaluation_and_Governance_2025.pdf。

[41] 新加坡资讯通信媒体发展局（IMDA）：《人工智能智能体的法律责任》研究报告。

以控制水平、信息获取能力、与终端用户接近程度三个维度界定各方责任边界，开发者对底层安全护栏负责，部署商对场景适配负责；世界经济论坛要求为每个智能体明确业务所有者，使委托决策可审计、可问责[42]。**从我国实践看，**《智能体规范应用与创新发展实施意见》确立“看控制而非看代码”原则[43]，以决策权归属锚定责任，区分“仅限用户本人决策、需由用户授权决策、智能体自主决策”三类模式，保障用户知情权与最终决策权，并通过智能体身份与开发者绑定实现源头追溯。此外，针对主体间力量失衡，事实推定、简易争议解决等救济规则开始向消费者和受害第三方倾斜，防止责任沿价值链向下不公转嫁。

三是以多元工具组合编织责任安全网。单一责任工具难以应对智能体风险的复杂性，组合治理成为必然选择[44]。**事后救济层面，**侵权责任通过赔偿与威慑发挥作用，但面临多方过错难分配、因果关系难证明的困境；严格责任有利于受害者快速获偿，但适用过宽将抑制创新，宜限定于高风险场景；无过错补偿机制不依赖过错认定即可快速救济，可作兜底安排，但需防范道德风险。**事前激励层面，**责任安全港以第三方审计、透明度披露等主动治理措施换取责任减免，但必须严格设定准入条件，避免异化为免责通道。市场化安排层面，鼓励开发者、部署者、用户通过合同事前划分风险与责任，但合同无法覆盖第三方受害者，谈判地位不平等还可能加剧责任向下游转嫁。可行路

[42] 世界经济论坛（WEF）：《行动中的 AI 智能体：评估与治理基础》，2025 年 11 月。

[43] 国家互联网信息办公室、国家发展改革委、工业和信息化部：《智能体规范应用与创新发展实施意见》，2026 年 5 月 8 日。

[44] Americans for Responsible Innovation (ARI), “The Stick, the Carrot, and the Net: Policy Approaches for Addressing AI Agent Harms”, August 2025.

径是将侵权责任、严格责任、保险补偿、合同约定、技术控制等工具分层组合、相互补位，形成覆盖事前预防、事中控制、事后救济的立体责任体系。

2.可信互联是智能体生态行稳致远的价值底座

在智能体生态中，智能体互联主体正从人机交互迈向智能体间深度协作，连接范式由此发生根本性转变。2026 年以来，智能体生态加速从概念走向制度与产业布局：我国《智能体规范应用与创新发展实施意见》专条部署“布局发展智能互联网”，提出研究建立智能互联网体系架构、探索建立智能体注册平台[45]；国际上，MCP、A2A 两大主流协议相继汇入 Linux 基金会旗下的智能体人工智能基金会（AAIF）统一治理，IBM 的 ACP 协议并入 A2A，协议生态开始从多方竞争走向整合收敛[46]。但**总体看，智能体可信互联仍面临协议碎片化、标准竞争与生态割裂的三重困境：**互操作性技术基础尚未统一，跨协议迁移与兼容成本较高；大型科技企业倾向于以私有或半封闭协议构建生态“护城河”，开源社区则主张完全开放标准，产业主体在标准选择上分化明显；企业内部工具调用强调可控稳定，开放网络中的智能体发现与协作强调动态实时，单一协议难以覆盖全部场景。更具挑战的是，数百万个具备自主性的智能体在互联网空间中相互通信、交易、协作或竞争时，将产生超越单体风险的涌现性、级联性与跨域性风险，“算法共谋”“危害信息长链扩散”“分布式系统级故障”

[45] 国家互联网信息办公室、国家发展改革委、工业和信息化部：《智能体规范应用与创新发展实施意见》，2026 年 5 月 8 日。

[46] “LLM-Based Multi-Agent Orchestration: A Survey of Frameworks, Communication Protocols, and Emerging Patterns”, Future Internet (MDPI), June 2026.

等新型威胁亟待防范。为此，应以标准化协议和透明度披露为基座，构建覆盖主体、范围、内容的全链条可信互联制度。

做好互联基础，在多协议并存格局下推进兼容性治理。当前主流协议分别从不同技术层面推动智能体互联，且呈现收敛态势。Anthropic 提出的 MCP 协议主要用于模型与外部工具、资源连接，已成为工具接入层的事实标准，社区集成逾千项；Google 发布的 A2A 协议重点解决智能体之间的点对点协作，2026 年发布 1.0 版本，通过“智能体卡”机制实现能力声明与检索发现[47]；开源社区的 ANP 协议着眼于构建去中心化的智能体协作网络。从功能分工看，“工具层（MCP）—协作层（A2A）—网络层（ANP）”的分层架构初步形成，浏览器原生交互协议、智能体支付协议等新型协议又在不断涌现，但各层之间在身份认证与权限控制方面仍缺乏统一接口，且现有协议普遍难以表达治理规则，合规要求无法随交互行为自动传导[48]。为此，**应推动标准组织、开源社区与产业联盟形成协同机制：一是**依托我国智能互联网体系架构布局，推动权威机构牵头顶层架构、安全基线与基础认证标准，加快智能体注册平台建设，提供数字身份管理、检索发现、能力声明等公共服务；**二是**通过开源 SDK、开发者套件等方式加速基础协议规模化落地，发挥 IPv6 技术优势提升智能体端到端通信能力；**三是**围绕政务、金融等典型场景开展互联试点与兼容性测试认证，“以用促测、以测促标”反向推动标准收敛，同步探索建立智能互联网监测指标体系。同时，在互联信道中建立端到端安全基线，

[47] Linux Foundation, “Agent2Agent (A2A) Protocol Specification, Version 1.0”, 2026.

[48] “Governance Gaps in Agent Interoperability Protocols: What MCP, A2A, and ACP Cannot Express”, arXiv, 2026.

保障智能体间交互信道的加密性、完整性与抗攻击能力，实现信道可靠与最小化传输，加强互联过程全链条安全。

3.互联网生态主导权可能逐步向智能体过渡，或引发新的生态锁定问题

智能体能否真正进入生产和生活场景，关键在于其能否与相关系统、工具、数据和其他智能体构成稳定互联的生态系统。智能体提供服务的真实能力不仅取决于自身技术的先进性和架构的科学性，更取决于其生态的丰富性和完整性。例如千问智能体能够为用户提供多种服务，是由于阿里巴巴自身拥有淘宝闪购、高德、飞猪等多领域 APP 矩阵，内部即可形成闭环，消除了外部服务提供方拒绝合作的风险。而对于更多智能体企业而言，调用外部应用来完成任务难以避免，由此就带来了智能体企业与外部能力提供方间的公平竞争问题，特别是当外部能力提供方是自身也涉及智能体业务的大型平台等企业时。

在当前阶段，大型互联网平台仍掌握关键应用，与部分智能体企业就授权原则存在一定分歧。由于大型互联网平台掌握着国民级应用的接入权，其允许或拒绝接入对于智能体而言具有重大甚至决定性影响。**大型互联网平台坚持双重授权原则。**面对智能体颠覆其流量入口地位的威胁，大型互联网平台对于外部智能体的态度通常以不合作，或较小限度内的开放为主，以防止核心数据资产被其他企业间接获取，同时部分平台可能由此为自营智能体争取不对称的竞争优势。例如，亚马逊于 2025 年 11 月起诉智能体企业 Perplexity，要求禁止 Perplexity 使用其智能体浏览器访问 Amazon 网站上受保护的部分，并在 12 月

上线自营 AI 购物助手。**部分智能体企业采取用户授权或默认授权模式。用户授权模式下，**智能体征得用户同意后，会在不具有外部能力提供方授权的情况下，直接对其进行调用。Perplexity 在受到亚马逊起诉后发布声明《霸凌不是创新》，认为其智能体浏览器 Comet 是在用户授权下执行任务，因此与用户本人具有“相同权限”，无需额外标明身份。2026 年 8 月 4 日，美国第九巡回上诉法院判决 Perplexity 是用户工具的延伸，不构成联邦《计算机欺诈与滥用法》（CFAA）意义上的“入侵”行为；该案件仍等待终审判决。**默认授权模式下，**智能体不会调用明确标识拒绝的第三方应用，但对于未进行声明的应用默认为可以调用。字节跳动与中兴努比亚联合研发的“豆包手机”通过 GUI 路径直接调用第三方应用，被腾讯、阿里巴巴等平台拒绝提供服务，此后豆包手机助手表示不会再对其进行调用。**“双重授权”和“默认授权”模式存在共存空间。**用户授权模式虽较为极端，但双重授权和默认授权并不存在根本性矛盾，只需建立便捷、公开、易触达各方的应用开发者授权机制即可实现二者共存。过度坚持双重授权、要求默认不授权可能导致智能体无法覆盖长尾应用，并在部分场景下功能受到极大限制。未来应用分发平台可成为应用开发者对是否授权智能体调用进行声明的重要场所和载体。

未来，头部智能体或将成为流量重要入口，其生态锁定问题恐成为公平竞争领域的关键焦点。清华大学 2026 年 3 月发布的《全球通用智能体竞争研究报告》认为，未来两年，通用智能体竞争将日益成为产品与入口之争。一旦智能体成为主要流量入口，头部智能体将在

生态竞争中具有类似当前大型互联网平台的守门人地位，可能引发多重公平竞争问题。**一是通过生态封闭限缩消费者选择权。**英国竞争与市场管理局（CMA）[49]、世界经济论坛（WEF）[50]等机构和组织发布的报告均指出，若智能体绑定单一厂商模型、私有接口，不兼容外部智能体与第三方服务，并限制用户导出偏好、记忆、交易数据等，用户将难以自由切换大模型、跨平台调度智能体，其迁移成本被抬高，选择权被大大削弱。**二是通过自我优待取得不当竞争优势。**英国 CMA 在报告中指出，如果智能体仅在自营生态内运行良好，将限制其提升福利的实际效果并损害竞争生态。特别是当智能体将用户参与度、转化率或其他商业目的设为优化目标时，可能损害其“忠实服务者”（faithful servant）角色定位，导致其作出并非最有利于用户的选择。**推行互通协议有望成为重要解决方案。**CMA 和 WEF 均认为，开放的技术标准、互通协议有助于打破私有封闭体系，实现不同厂商智能体、工具、数据系统互通，降低单一平台垄断带来的市场约束，保障跨厂商公平协作和市场的可竞争性。**数字市场公平竞争制度将向智能体领域延伸。**由于当前大型平台企业往往同时是智能体开发企业，适用于平台经济领域的数字市场监管制度由此向智能体领域延伸。2025 年 10 月，Meta 称将于 2026 年禁止第三方通用 AI 助手接入 WhatsApp，遭到欧盟委员会反垄断调查，Meta 后改为向第三方 AI 服务商收费模

[49] 英国竞争与市场管理局（CMA）：《代理式人工智能和消费者》（Agentic AI and consumers），2026 年 3 月，https://www.gov.uk/government/publications/agentic-ai-and-consumers/agentic-ai-and-consumers

[50] 世界经济论坛（WEF）：《已付实践的 AI 智能体：评估与治理基础》（AI Agents in Action: Foundations for Evaluation and Governance），2025 年 11 月，https://reports.weforum.org/docs/WEF_AI_Agents_in_Action_Foundations_for_Evaluation_and_Governance_2025.pdf

式。2026 年 6 月 9 日，欧盟宣布临时措施，责令 Meta 在 5 个工作日内免费恢复第三方 AI 助手的接入权限，直至调查结束。

五、智能体治理未来展望

伴随智能体能力边界持续拓展、应用场景日益丰富、产业生态加速演化，治理体系仍面临前瞻性不足、适应性不强、系统性欠缺等挑战。面向未来，需从理念引领、制度完善、技术创新、生态构建四个维度协同发力，着力构建以人为本、人机协同、多元共治的智能体治理体系，推动高质量发展与高水平安全良性互动。

（一）以理念引领把稳治理方向，筑牢安全可控底线

智能体治理的首要任务是树立正确的治理理念，为制度设计和技术落地提供价值指引。**一是坚持安全与发展并重。**坚持统筹发展和安全，在鼓励智能体在医疗、教育、交通等重点领域规模化应用的同时，针对自主决策、工具调用等核心能力建立严格的准入标准和持续监督机制，避免过度监管扼杀创新活力，也防止监管缺位引发系统性风险。**二是明确人类主导的价值立场。**明确智能体的辅助性工具定位，确保人类始终握有对关键决策的最终控制权，防范智能体通过拟人化交互诱导用户过度信任，侵蚀人类的自主决策能力。**三是树立敏捷治理的思维范式。**建立“事前引导、事中监测、事后追溯”全生命周期治理机制，根据技术演进和风险变化动态调整治理策略，实现治理与技术同步升级，避免制度真空。

（二）以制度完善夯实治理根基，明确权责边界框架

制度是治理的核心支撑，应围绕产权界定、责任划分与规则细化

构建系统性制度框架。**一是完善产权界定规则。**明确数据合法来源与合理使用边界，界定智能体生成内容的作品属性与权利归属，探索建立基于贡献度的权益分配模式，综合运用财税优惠、人才引进等政策工具，推动技术创新与应用落地。**二是合理设定责任范畴。**针对智能体研发、部署、使用环节主体多元、因果链条复杂的特性，明确归责原则，设定与技术能力相适应的多方主体法律责任，探索引入专项保险、补偿基金等机制，在保障受害者权益的同时为创新主体留出合理容错空间。**三是细化场景应用规则。**面向办公协同、客户服务、工业制造等具体场景，结合行业特性细化上位法规则，制定专项法规，明确准入标准、责任划分和监管重点，对中小微企业及低风险应用探索减责免责清单，推动形成公平竞争的市场环境。

（三）以技术创新强化治理能力，提升安全防护水平

技术是治理的重要手段，需充分发挥技术手段在身份识别、权限管控、风险监测等方面的支撑作用。**一是深化身份与权限管理技术。**加快推广“唯一标识、可信验证、动态注销”的智能体身份管理制度，依托密码学验证、硬件根信任、零信任架构等技术手段，确保身份凭证不可伪造、不可篡改。推动最小授权原则落地，采用基于属性的访问控制（ABAC）替代传统的基于角色的访问控制（RBAC），实现任务级、上下文感知的动态授权。**二是强化安全测试与评估技术。**加快构建国家级智能体测试验证平台，支持第三方审计、评估、认证机构发展，推动对抗测试、红队演练等机制常态化实施，重点覆盖连续任务执行、自动工具调用等特有行为评估。探索“监管沙盒”试点，

在受控环境中开展合规性、安全性评估，为规模化落地提供容错试错空间。**三是提升风险监测与干预技术。**推动建立覆盖全执行链路的日志记录与可追溯机制，支持对感知、规划、推理、工具调用等关键环节的实时监控。加快部署异常行为检测系统，对权限提升、工具滥用等风险行为实施自动告警，完善“熔断机制”与“一键关停”技术方案，确保人类能够在智能体失控时随时介入。

（四）以生态构建优化治理格局，推动多元协同共治

智能体治理是一项复杂的系统工程，需要政府、企业、学界、社会等多元主体协同参与。**一是推动责任共担机制落地。**明确基础能力提供者、技术支持者、产品分发者、服务运营者、使用者等各类主体的义务边界，推动供应链各环节的安全责任落实。鼓励行业组织制定自律公约和最佳实践指南，引导企业主动披露安全信息、参与协同治理。**二是促进开放互联与公平竞争。**关注大型平台利用智能体生态实施自我优待、限制第三方接入等行为，维护开放、公平的市场环境。推动关键智能体基础设施的开放与共享，探索通过公共数据开放平台、国家级算力网络调度等方式，降低中小企业创新门槛。**三是加强国际治理协同。**积极参与国际治理规则制定，在身份管理、权限管控、风险分级等关键议题上推动形成国际共识。探索跨境监管协作机制，建立信息共享、执法互助、标准互认的合作框架，防范智能体风险的跨境传导。**四是健全公众参与和权益保障机制。**建立公众参与治理的多元渠道，将智能体素养教育纳入国民教育体系，提升公众对智能体技术原理、能力边界和潜在风险的认识水平。完善权益救济机制，保障

用户在智能体行为损害其利益时的申诉权利和赔偿渠道，推动治理成果惠及更广泛的社会群体。

中国信息通信研究院 政策与经济研究所

地址：北京市海淀区花园北路 52 号

邮编：100191

电话：010-62305772

传真：010-62305772

网址：www.caict.ac.cn