智能体发展与治理研究报告

AI 善治学术工作组

2025 年 10 月 17 日

中国·重庆

编 委 会

主编/策划　AI善治学术工作组

张凌寒	中国政法大学人工智能法研究院教授、院长
杨建军	西北政法大学法治学院教授、《法律科学》主编
程　莹	中国信通院政策与经济研究所高级工程师
赵精武	北京航空航天大学法学院副教授、院长助理
韩旭至	华东政法大学数字法治研究院副教授、副院长
郑志峰	西南政法大学科技法学研究院教授、副院长
徐小奔	中南财经政法大学知识产权学院教授、副院长

编写团队

于晓洋	中国人民大学法学院博士后
徐坤杉	中国政法大学数据法治研究院博士研究生
薛少雄	中南财经政法大学知识产权学院博士研究生
杨长元	西南政法大学人工智能法学院博士研究生
刘芮池	中国人民大学法学院硕士研究生
高　华	中国人民大学法学院硕士研究生

特别鸣谢

特别鸣谢，为此报告提供特别指导的各家企业，

感谢各位对报告编制的大力支持！

目 录

第一章 全球智能体发展态势

一、智能体概念日渐明晰，体现三大核心特征

（一）智能体的概念与特征

当前，智能体（AI Agent）的概念界定呈现多维视角，不同研究机构基于各自关注重点提出了差异化定义：从技术实现视角看，IBM 将其定义为“通过程序化工作流程和工具调用实现自动化任务执行的智能系统”，强调了其在任务分解、流程规划和工具适配方面的技术特征。从人机交互视角看，Salesforce 公司提出 AI Agent 是“能够自主理解并响应客户需求的智能服务系统”，突出了其在自然语言处理、意图识别和服务交付等方面的交互能力。从系统架构视角看，国际数据公司（IDC）将其界定为“以大语言模型为核心，具备环境感知、决策推理和行动执行闭环能力的自主系统”，着重描述了其“感知－思考－行动”的三位一体架构。从功能特性视角看，复旦大学 NLP 团队提出：“AI Agent 是能够自主感知环境信息、进行决策推理并执行目标导向行动的智能系统或实体”（如图 1）。

在国内企业的实践中，智能体的概念逐渐清晰，并呈现出多元化的发展方向。MiniMax 认为，当前智能体位于 OpenAI 提出的 AI 进化等级体系中的 L3 级别，即“能采取行动的代理型 AI 系统”，是 Chatbot 向具备任务执行能力的 Agent 系统演化的重要阶段。百度则将智能体理解为具备双重系统的结构：系统 1 具备推理和记忆能力，可调用外部数据库完成任务；系统 2 具有理解、规划和反思等高级认知能力，二者协同响应用户需求。

在其技术路径中，智能体被视为文心大模型的子集，并非独立模块。

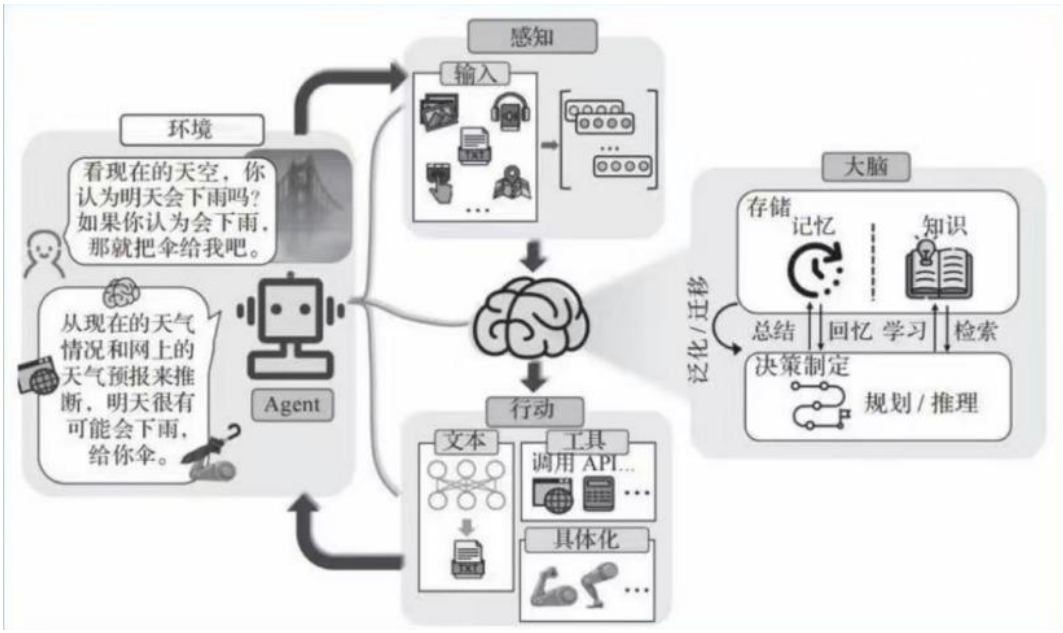

图 1：LLM-based Agent 的工作原理（图片来源：复旦大学 NLP 团队）

综合现有研究成果，本报告将 AI Agent 界定为：以人工智能技术（特别是大语言模型）为支撑，具有环境感知与语义理解双重能力，可基于预设目标自主完成任务分解、推理决策和工具调用的数字化智能实体。就实现形态而言，智能体既可体现为软件程序等虚拟存在，也可具象化为机器人等物理实体。

总体来看，智能体主要具备以下三个核心特征：

一是自主性。智能体能够在无人干预的情况下，独立感知环境、自主决策并执行任务。其具备自我管理与驱动能力，能够围绕设定目标独立运行。例如，在智能家居系统中，智能体可持续监测室内温度与湿度，自主调控空调与加湿器，无需人工干预即可营造舒适的居住环境。

二是交互性。智能体能够与人类用户、其他智能体或环境实体进行有

效的信息交流与协作。其交互方式不限于简单的“指令－响应”，而是具备上下文理解、多模态信息处理与社会化协同能力。例如，自动驾驶系统可通过与交通信号灯、周围车辆和行人持续交互，确保安全行驶。

三是学习性。智能体能够基于历史经验、交互数据与环境反馈，借助内置机器学习算法不断优化其决策模型与行为策略，使智能体的性能持续改进提升。例如，虚拟助手通过不断学习用户的表达习惯，能够更准确地理解口语化、模糊的指令，并作出精准回应。

（二）智能体的工作原理

在人工智能技术迭代发展的浪潮中，基于大语言模型的 AI 智能体（LLM-based Agent）已成为当前研究与应用的前沿热点。这类智能体系统由三个高度协同的模块构成：感知端（Perception）、控制端（Brain）和行动端（Action）。感知端作为环境信息采集门户，采用多模态数据融合框架，不仅具备自然语言处理能力，更整合了计算机视觉、语音识别等多项感知技术，实现视觉、听觉等多维度环境信息的实时采集与预处理。控制端作为系统决策中枢，依托大语言模型强大的知识存储与推理能力，不仅能实现知识图谱的动态构建与更新，更能执行复杂的目标分解、任务规划及策略生成，其特有的迁移学习特性赋予系统持续自我优化的能力。行动端作为环境交互实现单元，通过 API 工具调用系统与具身智能系统的双重实现机制，既可以在数字化场景中完成精密操作，也能通过机器人终端在物理世界实现实体交互。这三个模块通过信息流与控制流的闭环连接，共同构成了完整的“认知－决策－执行”循环体系。以自动驾驶系统为例，

智能体首先需要感知周围的交通情况和道路状况等信息，然后根据感知的信息决定是否加速、减速或转弯等，最后根据决策执行相应的行动，包括控制汽车的加速器、刹车和方向盘等。这种闭环运行机制充分展现了智能体最本质的特征——环境自适应性，即在无人为干预的前提下，通过实时环境感知与自主决策实现复杂场景下的智能行为调控。

（三）智能体的分类

一是基于系统规模的分类，可以分为单个智能体和多智能体。单智能体独立运作来实现特定目标，适用于边界明确的简单任务场景。其利用外部工具和资源来完成任务，从而增强在不同环境中的功能。多智能体指通过协作或竞争来实现共同目标或各自目标的多个智能体系统，可以综合调动、利用各个智能体的不同能力和角色来实现复杂任务的处理，每个智能体都可以拥有最适合其需求的基础模型。多智能体系统可以在互动场景中模拟人类行为，如人际沟通。

二是基于决策机制的分类，可以分为简单反射型智能体和基于模型的反射型智能体。简单反射型智能体仅能根据当前输入信息执行预设的规则动作，无法存储或处理历史数据。它们通常采用“如果...那么...”的条件触发机制，适用于标准化程度高、变化少的业务场景。虽然其逻辑简单、响应迅速，但由于缺乏学习能力和环境理解深度，在复杂度高的任务中表现受限。基于模型的反射型智能体通过内置的环境模型来构建对世界的认知框架，能够推理当前未直接观测到的信息。它们比简单反射型智能体具有更强的适应性，可应用于需要预测分析的场景，如工业设备故障预警或

金融风险预估。这种模型驱动的特性使其能够在信息不完整时做出更合理的决策。

三是基于目标特征的分类，可以分为效用驱动型智能体和目标导向型智能体。效用驱动型智能体通过可量化的效用函数来评估不同决策方案的价值，最终选择综合收益最高的行动路径。这种基于概率和统计的决策机制使其特别适合资源优化类的任务，如物流配送规划或电力调度。通过精确量化各项决策指标，它能够在多重约束条件下找出最优平衡点。目标导向型智能体的核心设计围绕特定目标的达成而非简单的即时响应，具备任务分解和路径规划能力。它们可以在复杂环境中通过多步操作策略性地达成目标，例如自动驾驶车辆的导航系统。其优势在于不仅能执行任务，还能根据环境变化动态调整实现路径。

二、智能体行业高速发展，初现通用智能雏形

从技术路径来看，智能体正从“感知－规划－记忆－行动”四维架构向更高阶的自主智能演进。多模态大模型、检索增强生成、大小模型协同等技术加速融合，不断提高智能体的感知能力、学习能力与任务适应能力。例如，2025 年 4 月，Monica 公司发布核心产品 Manus，其作为全球首款通用 AI 代理，能够自主拆解任务、规划步骤并调用工具执行。未来，智能体的演进将呈现双重驱动：在模型层面，借助小模型与 MoE 架构实现轻量化蜕变；在算力层面，为满足实时响应与降低带宽等需求，从集中式向“云－边－端”协同范式重构，为智能体技术的进一步跃升奠定基础。

从应用场景来看，智能体已从概念验证走向规模化落地，正在深刻重

塑金融、医疗、工业、政务、教育、文旅等各行各业的运作范式。其中，智能体在智能客服场景的渗透率高达70%、在数据分析场景渗透率达到60%，展现较高的应用成熟度，同时蕴含研发、营销等场景的爆发点。[1]例如，沃尔玛利用智能体实现缺货风险的动态预测，将响应时间大幅压缩至15分钟；[2]一汽丰田通过部署腾讯云智能体，将客服独立解决率从37%显著提升至84%。[3]智能体正在全行业掀起智能化变革浪潮。随着技术的持续成熟，智能体正从“效率工具”向“价值创造者”演进，成为推动产业数字化升级的核心驱动力。

从产业布局来看，智能体呈现出市场规模爆发式增长、行业生态双轨并行的格局。一方面，智能体产业规模增长迅速，2023年中国AI Agent市场规模为554亿元，预计2030年将增长至8250亿元。[4]另一方面，智能体行业生态形成“通用平台－垂直场景”双轨并行发展的格局。在通用平台型智能体层面，海外头部厂商在底层模型架构与开发生态争取行业高地，Google推出的Gemini CLI热度持续攀升，OpenAI推出Operator等多种智能体产品。在垂直场景智能体层面，国内厂商深耕具体应用领域，迈富时、美洽、玄武云、神州云动、蓝凌等厂商深入行业痛点，打造销售、客服、快消、CRM、办公等场景化解决方案。

[1] 第一新声智库：《2025年中国企业级AI Agent应用实践研究报告》
[2] 甲子光年：《中国AI Agent行业研究报告（二）》
[3] 同1
[4] 头豹研究院：《2024年中国AI Agent行业研究：创新驱动，智能技术革新》

三、智能体驱动社会变革，助力实现降本增效

（一）重构生产模式：从流程执行到自主运营

AI Agent通过其自主性、记忆能力、权限管理和工具使用，彻底改变了生产力的范式。微软认为，AI Agent不仅是一种为人们获取更多价值的工具，还将彻底改变工作完成的方式。具体来看，一是通过自主性重塑工作模式。智能体能够全天候工作，处理从简单到复杂的多步骤任务。例如，审查客户退货、优化供应链流程、生成销售报告或为技术人员提供实时指导。二是通过记忆提升连续性。在记忆方面，AI代理不仅能够执行任务，还能通过记忆功能提供上下文连续性，避免重复劳动。三是通过权限管理确保安全合规。在权限方面，它可以通过权限管理确保安全访问企业数据。四是通过工具调用优化工作流程。在工具使用方面，可以通过工具集成直接采取行动。这种能力使员工从繁琐的日常事务中解放出来，专注于更具战略性和创造性的工作，同时推动企业整体效率和创新能力的提升。2025年3月，OpenAI推出了专门用于构建智能体的Responses API和Agents SDK，内置网页浏览、文件检索、电脑操作等工具，显著简化了开发人员构建智能体应用的流程。

（二）重塑协作范式：从线性作业到全域并进

AI Agent通过其卓越的并行处理能力、持续服务特性、智能分析功能和精准执行机制，正在重构人机协作的效率体系。这种变革主要体现在以下四个方面：一是并行处理加速任务响应。智能体通过多线程机制可同步处理多项任务，显著缩短响应时长并提升服务吞吐量。此外，系统支持基

于用户价值的优先级调度算法，在确保服务质量的前提下，优先响应高价值客户的紧急需求，实现服务效率与质量的双重优化。二是7×24持续服务优化时效性。区别于人工服务的时间限制，智能体具备全年无休的持续服务能力。这种全天候特性有效强化了用户粘性与品牌忠诚度。三是智能分析驱动运营提效。智能体在交互过程中持续生成的用户行为数据（包括咨询热点图谱、服务路径转化率、需求波动周期等），为企业提供了精细化的运营洞察。四是精准输出减少人为误差。智能体可对用户提问做出一致而准确的回复，从而降低出错风险，确保用户获得可靠的信息。它们主要通过代理循环和类似人类的思考过程提高回复的准确性。有研究表明，金融投研智能体缩短80%尽调周期、医疗诊断Agent误差率降至2.3%、工业运维预测准确率达91%等。

（三）革新交互方式：从指令操作到自然对话

交互方式是指人类与智能体或智能体之间进行信息传递、任务执行和社会协作的具体模式。随着智能体技术的突破性发展，交互方式正经历着从机械式指令向类人化协作的范式跃迁，这种变革不仅重新划定了人机协作的能力边界，更在根本上推动着社会数字化转型的进程。当前变革主要体现在以下三个维度。

一是任务交互的革命性简化。传统的人机交互需要用户将复杂需求拆解为多步操作，而智能体能够直接理解并执行高级任务，让交互真正实现“所想即所得”。例如，用户只需提出“策划一场跨境电商营销活动”这样的复合需求，智能体即可自主完成市场分析、创意生成、渠道投放等全

流程任务，从而显著降低使用门槛，提升效率。二是情感化与拟人交互的突破。智能体可以模拟类似人类的社交行为，例如建立关系和分享信息。2024 年底，无界方舟（AutoArk）推出全球首款个人基础智能体 Arki One，其具备完善的情绪系统，支持 21 种语言的流畅交流，并能在百毫秒级精准驱动数字人或智能硬件的表情、动作与语音协同，使交互体验更趋近于人与人之间的沟通。三是社会化行为与自主协作的涌现。智能体不仅能与人类交互，还能在多智能体之间形成动态协作网络，自发产生复杂的社交行为。2023 年，斯坦福研究人员成功地构建了一个名为 Smallville 的虚拟小镇，25 个智能体在数字社区中自发形成了复杂的社会关系网络。它们既扮演着园丁、作家等职业角色，又能像真实人类一样参与社交活动、结交朋友，甚至自发组织派对。每个智能体都具有独特的人格特质和行为模式，使整个数字社区呈现出近似人类社会的动态演化特性。

第二章 智能体应用场景：纵深拓展，走深向实

一、智能医疗：全周期管理实现精准化干预治疗

医疗健康领域中，人工智能大模型可以实现的是医学知识的获取与文本输出，智能体可以实现的则是代理医疗实践。以往，用户往往依赖自身经验与医疗大模型进行即时对话，进而了解相关病症知识。在此情形下，即使模型配备了专业度较高的医学数据库，由于用户普遍缺乏医学背景，专业性不足，难以正确描述病症、引导对话，甚至对模型输出的答案产生错误理解，引发不必要的误解与过度焦虑。医疗智能体则可以通过自主调用、分析、决策，深度利用用户数据，必要时结合对话交互，提供“量体裁衣式”的健康管理，并提供全周期的科学化预测服务。

例如，清华大学成立的 Agent Hospital 为业内提供了智能体医疗的新型范式参考，其中包括模拟构建和代理进化。[5]医生智能体和护士智能体均可以为患者提供相应的服务，若成功治愈，该案例将被记录下来，以供后续参考。反之，则会开始新的治疗周期。医生智能体在业余时间还会进行自我提升，阅读医学书籍来巩固自身专业知识和技能。清华团队对比了医学智能体进化前后在各个疾病上的诊断准确率，发现均有大幅提升，验证了其自主进化的有效性（如图 2）。又如，四川大学华西医院与华为联合研发的睿兵 Agent 可以根据患者实际情况，提醒患者重点关注事项、按时

[5] Li J, Lai Y, Li W, et al. Agent hospital: A simulacrum of hospital with evolvable medical agents[J]. arXiv preprint arXiv:2405.02957, 2024.

到院复查复诊，[6]这也显著优化了患者的就医体验。以往智能医疗开始于患者对病症有所感知的时刻，而智能体则可以将这一干预时点大幅提前。在技术成熟的情况下，智能体可以在用户尚未察觉异常之前就主动触发医疗服务机制，提醒患者及时就诊。从数量上看，医生智能体能够治疗的患者数量远高于人类医生所能治疗的患者数量；从质量上看，医生智能体还可以长期保持迭代，不断提高医疗水平。换言之，医生智能体不仅可以提供更加精细化、精准化的治疗方案，还可以促进实现“一人一策”乃至“一人千策”的医疗升级。

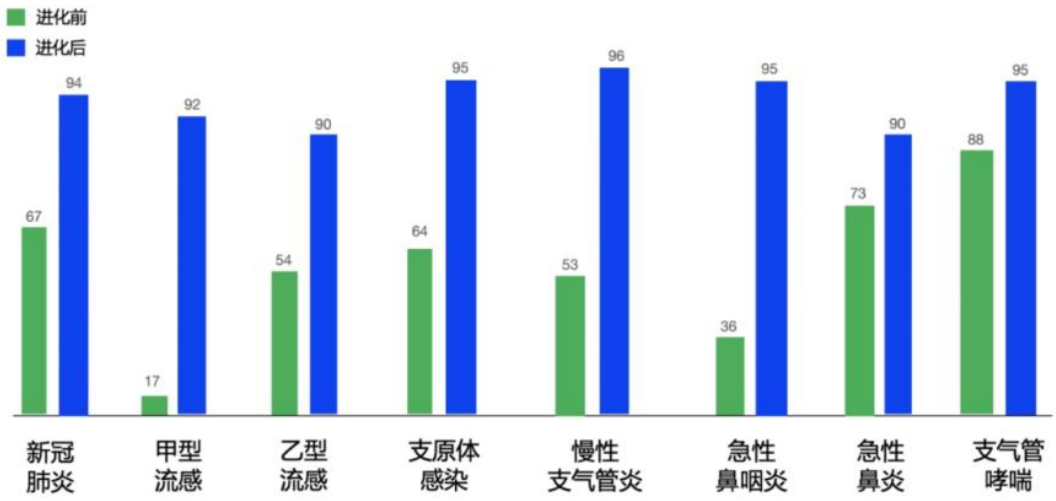

图 2：智能体进化前后不同疾病的诊断表现（图片来源：清华大学智能产业研究院）

此外，通过调适和设定，智能体也可以赋能医疗服务更具温度，帮助用户及时高效匹配符合需求的医疗资源。例如，智能体可以综合分析患者的病情、收入预算、地理位置远近、医保受惠情况等特点以及医院的擅长

[6] 四川健康传播：《专注消化领域 华西医院胡兵教授团队联合华为等发布医学 AI 智能体“睿兵 Agent”》

领域、价格定位、基础设施等因素，帮助患者筛选合适就诊的医院以及符合要求的医生。一方面，智能体弥补了患者个人能力的缺陷，提高就医效率；另一方面，智能体协同医疗资源的合理分配，提高医疗体系的运行效率。推动智能医疗高质量发展，实现有温度的智能服务。

二、智能交通：一体化调度打造人车路协同生态

交通领域中，智能体推动了个人驾驶和公共交通两个层面的系统性变革。在个人驾驶层面，智能体赋能自动驾驶突破感知边界，迈向“无感”交互的全新阶段。2025 年 3 月智己汽车携手阿里斑马智行发布了 IM AIOS，无需触控，仅需语音交流即可完成操作，将生态服务以 AI Agent 形式落地，并且联合外卖 APP、票务 APP 打造点餐、购票 Agent。[7]例如，用户只需说“点个微辣麻辣烫”，车内智能体就能精准识别语音，并进行无感支付完成下单，结合导航信息和历史数据选定餐厅，实现“人到、车到、餐到”。未来，车控智能体还可以实现与各个服务场景的智能串联，如与公共交通智能体系统进行互联互通，进一步降低事故发生率，维持更加长期、平稳、安全的驾驶体验，推动智能交通向高水平安全发展。

在公共交通层面，智能体促使其智能化转型，优化路线选择、时间调整等方面的决策，实现自动调度、流量优化和安全防御升级。整体来看，以智能城市为例，多智能体系统（MAS）能够实时管理交通流量，通过车对万物（V2X）通信，使车辆能够与其他车辆、行人和道路基础设施进行

[7] IM 智己汽车：《一图看懂 IM AIOS 发布会，智己携手阿里系 AI 进入 No Touch & No App 新时代！》

交互。[8]在重大突发事件或极端天气等紧急情况下，智能体可以通过多模态协同，整合区域内的交通网络资源与运力，提升路网应急响应能力，保障公共交通系统运行更安全、更高效。具体来看，智能体可以提高主动智能信号调节、实时路况精准感知与预测等方面的服务性能。例如，智能体赋能交通信号灯具备主动调节的能力。同时，借助协同数据库，公共交通智能体可以进行实时监控和动态调整，实现自主规划路线、发车分配、站点管理等调度任务。例如，中科视语智能网联云控平台（如图 3）以车辆、道路、环境的动态数据为核心，结合多源感知，实现协同感知、决策、控制，推动车路协同管控与调度。

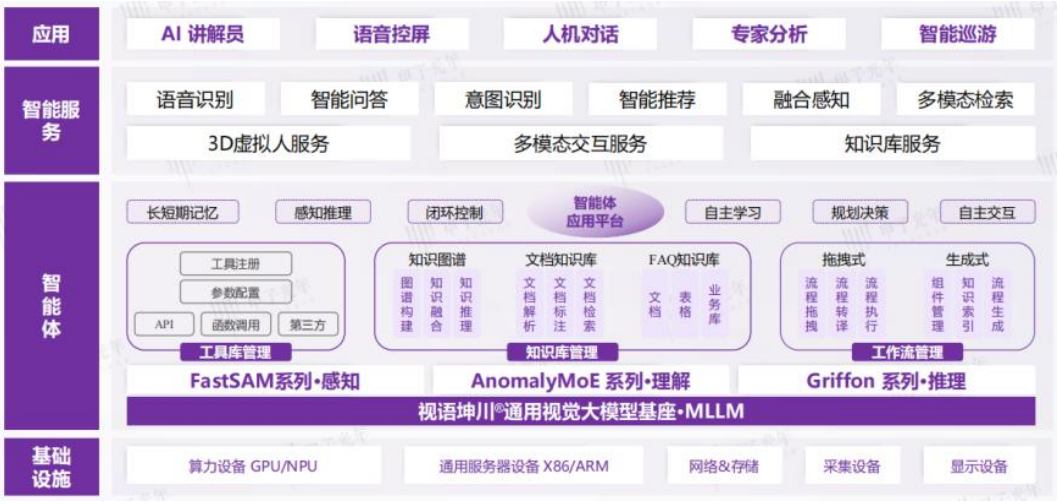

图 3：中科视语智能网联云控平台（图片来源：甲子光年）

此外，公共交通智能体也可以综合分析驾驶员数据、车辆行驶数据以及线路数据等交通关键影响因素的数据，提高预测公交系统故障、驾驶员疲劳驾驶、线路风险等交通问题的精准度，并以高度自主决策和执行能力

[8] World Economic Forum：《Navigating the AI Frontier: A Primer on the Evolution and Impact of AI Agents》

实现突发事件及时响应、自主求助和安全系统升级。这有助于形成地面公交、地下地铁以及地上单车的多方联动服务生态，实现“定点接人、按时到站”，有效提高了公共交通的服务效率和质量。

三、智能教育：师生双向赋能实现“因材施教”

教育领域中，教育智能体能够同时面向师生双主体，以师生需求为导向持续拓展应用，针对不同场景、角色实现智能化精准服务。这不只是智能教育的形式发生改变，更是智能教育质量的一次跃迁。例如，内蒙古医科大学开发的教育智能体“精诚智教”（如图 4）展现了细化的教育智能体的生态架构。

图 4：精诚智教 本科教育综合智能体生态架构（图片来源：内蒙古医科大学）

又如，基于 Deep Seek- R1 本地化部署的“浙大先生”智能体平台，赋能教学、科研、治理等教育细分场景，面向全国师生群体提供前沿服务，推进智能教育深度应用。未来，教育智能体可以实现资源调动配置，为不同地区的学生配置相适应的资源并制定进阶方案，拓展教育资源共享的边界，在一定程度上促进了教育公平发展。在此基础上，教育数据解析与利用更加深化，能够有效驱动智能教育的质量进一步提升，引领智能教育体系和能力真正走向高质量发展。

面向学生，教育智能体推动智能教育从“解答工具”走向“智能导师”。相较于原有只能通过拍照或复制习题向智能教育 APP 提问的方式，教育智能体则可以全程参与学生的学习过程，支持学生实现个性化学习和跨学科主题探索，能够实时监测学习进展，精准捕捉当前学习障碍，通过互动来持续跟进学习情况，主动提供解析或相似习题练习，制定未来的学习目标，从而实现针对性的因材施教。面向教师，教育智能体的应用既可以提高教学管理效率，也可以提升科研效率。一方面，教师可以使用智能体助教，覆盖教学全流程，自建开发智能体提升教学管理效率；另一方面，教师还可以通过智能体获取选题推荐、实验模拟、申报优化、数字化盲审及进度管理等教改项目全流程智能支持。可以看出，教育智能体呈现出了“更自主、更适应、更高质”的特点。

四、智能物流：全链路协同实现资源自动化部署

物流领域中，物流智能体是融合 AI、物联网等技术的智能化实体，覆盖仓储拣货、安全驾驶等多个场景。[9]智能体赋能物流可以实现全过程自动化，推动运输系统“自选司机、自动找路、自主运输、自调库存”（如图 5）。当前，阿里菜鸟、京东物流、东航物流等国内头部物流企业，亚马逊、德国邮政敦豪等国外物流服务提供商，均已推进 Agent 在物流运输链之中的应用。

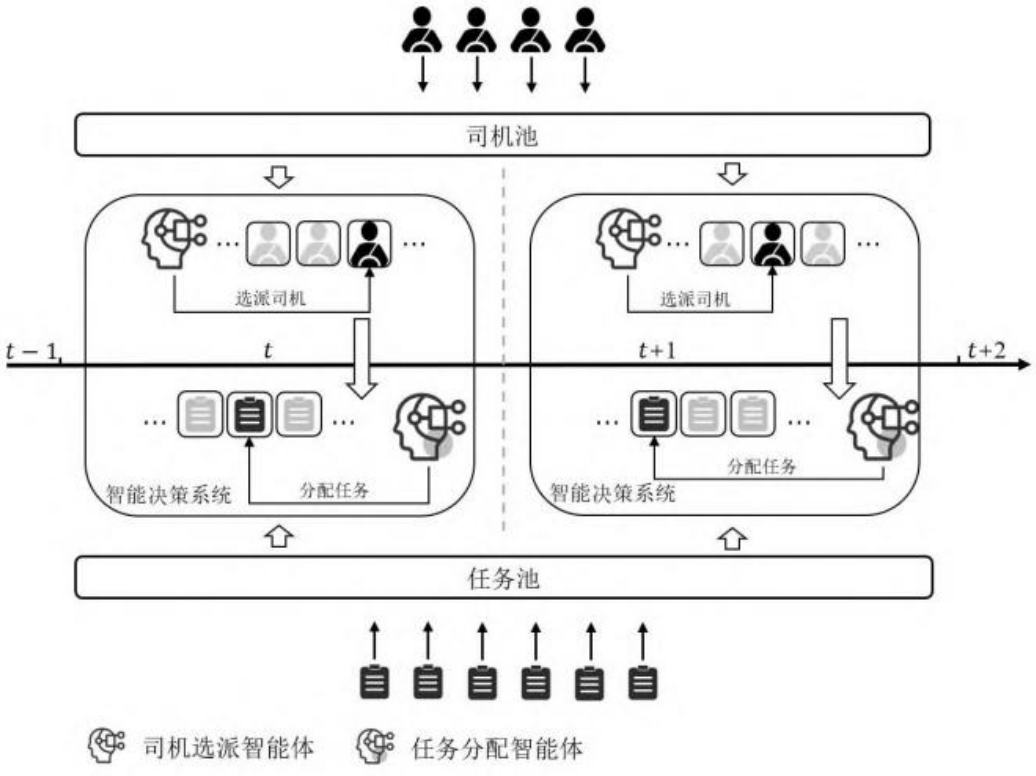

图 5：双智能体协作任务分配系统示意图（图片来源：《双智能体协作学习的众包物流任务分配模型》）

[9] 九州通医药集团：《AI 智能体+物流 | 九州通 AI 创新应用之物流篇》

首先，智能体通过融合道路通行、实时路况、管辖机构、气象条件等数据，能够持续优化运输路线并及时反馈到运输人员，提高整体效率并保障过程安全。例如，在矿山内部运输中，智能体可以根据实时路况和车辆负载情况，为运输车辆规划最优路线，减少运输时间和能耗。[10]其次，智能调度系统可以帮助实现车货精准匹配以及人路最优协同，显著降低车辆空载率和物流运输成本，为智能物流的可持续发展提供了有力支持。例如，在 72 小时实测案例中，实验组车辆空驶率降低 23.7%，紧急订单响应速度提升 41.2%，碳排放减少 18.9kg/百公里。[11]最后，智能体能够深度融合分析承运人、客户、仓库、货物以及订单等多维数据，建立高效协同机制，自动修正库存方案，提高仓库管理的安全性，降低综合管理成本。

具体而言，物流智能体的显著优势是能够大幅降低决策和环节调整成本，这是普通人工智能按照固定程序执行难以实现的。一方面，物流智能体能够结合不同地域的订单数量规模，动态调整跨区域的运输资源的分配，提升整体物流运输效率，实现异域运输资源利用最大化。另一方面，物流智能体还会深度引入地图和路线规划系统，结合本地区路况特点、配送时段的交通情况、距离远近、配送人员设备与身体状态等多种要素，实时更新当下时段内用时最短的配送方案，实现高效末端配送，优化同城配送流程。此外，智能体可以降低报价过程中的人力成本。结合自身运输网络的布局特点，物流智能体可以进一步精准生成性价比最高的最优报价方案，并且随着运价波动而实时更新，实现智能调价。

[10] IntelMining 智能矿业：《AI Agent 行业研究：矿山行业迎来智能体时代》
[11] Xu L, Mak S, Schoepf S, et al. Multi-agent digital twinning for collaborative logistics: Framework and implementation[J]. Journal of Industrial Information Integration, 2025, 45: 100799.

五、智能金融：形成风控闭环与 BC 端多维保障

金融领域中，智能体的深度赋能可以满足多样化数据、专业知识、动态决策等需求，金融机构服务效率能够得到显著提升，运营成本也可以有所降低，智能体金融风控、市场交易、客服运营等方面有着广泛应用空间。

风控方面，世界经济论坛指出智能体可以增强欺诈检测的能力。[12]金融智能体可以构建“实时监测+预测预警+自主迭代”的闭环工作模式，实现信用评估革新、复杂欺诈识别、合规自动化的全流程防护。通过对市场内交易数据流和用户行为的实时监测，金融智能体在内部搭建起庞大的风险感知网络，能够适时进行合规管理，加强了面向风险的抵御能力。例如，通付盾推出了风控智能体“神烦狗”能够自动捕获绕过规则的异常样本，动态分析生成新策略建议。[13]

交易方面，智能体能够根据市场变化情况和预测趋势，结合客户的风险偏好数据，提供定制化的资产配置建议，自主调节投资组合中的资产比例，实现风险和收益的均衡。例如，智能体平台 Unique 在资产管理方面处于领先地位，客户包括 Pictet Group、UBP、SIX、LGT 等知名金融机构，管理资产规模超过 2.3 万亿美元。[14]又如，智能体驱动生成金融数据分析报告的“TRACE”框架（如图 6），显著改善了传统金融数据分析在应对高时效性、强交互性的业务时出现的延迟与滞后。

[12] World Economic Forum：《The rise of ‘AI agents’: What they are and how to manage the risks》
[13] 通付盾：《通付盾风控智能体（RiskAgent）: 神烦狗（DOGE）》
[14] 每时 AI：《金融 AI Agent 平台 Unique，获 3000 万美元融资》

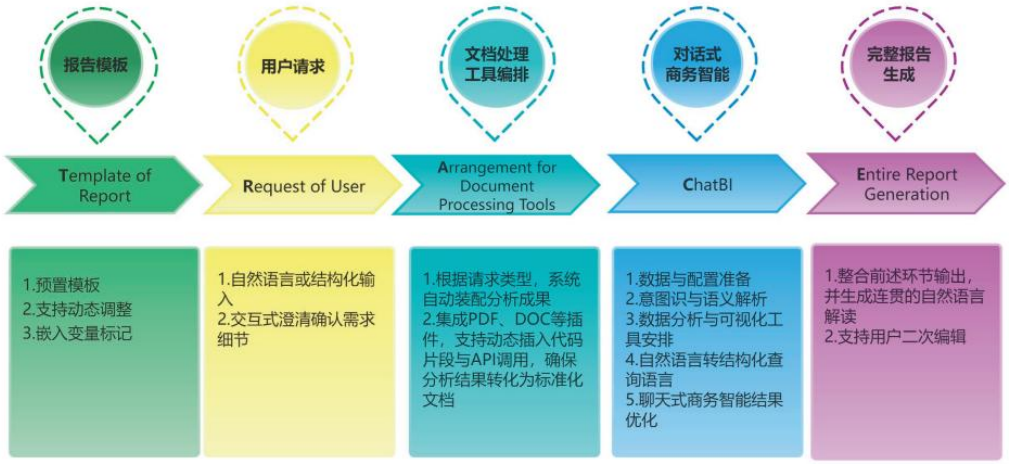

图 6：面向金融数据分析报告生成的“TRACE”框架（图片来源：金融电子化）

运营方面，金融智能体解析用户指令和请求的能力更强，通过多模态情感交互和上下文语义理解，可以为用户提供更加贴心的服务和更精准的解答。例如，蚂蚁金服的智能客服系统通过自然语言处理技术，能够精准理解客户的语义，快速回答常见问题，处理比例高达 80%以上。[15]凭借高感知力，金融智能体还可以通过情感分析识别用户的情绪状态，当用户表现出焦虑或不满时，智能体能够及时安抚用户情绪。这提高了金融机构受理咨询、办理业务以及解决问题的效率。

六、智能制造：自主决策优化推动提质降本增效

制造业领域中，智能体的应用可以更大程度地减少人工干预环节，推动制造全过程自动化，提高制造业生产效能并降低生产成本，实现“更短周期、更少人力、更多成果”。Deloitte 的一项分析预测显示，到 2027 年采用生成式 AI 的公司中将有一半推出“智能体”试点，用以执行复杂任

[15] 中关村互联网金融研究院：《智驱金融新图景：AI Agent 如何打通应用 “最后一公里”》

务，且几乎不需要人工监督。[16]目前，已有研究构建出制造智能体的技术架构全景（如图 7），包含感知层、决策层和协同层三个层面。数据显示，工业智能体的规模化应用可使制造业综合效率提升 20%～40%，2025 年的市场规模预计突破 5000 亿元。[17]实践中，美的集团已经投入建设工厂智能体，其中包括计划智能体、设备智能体、品质智能体等多个具体应用。[18]

图 7：智能体技术架构全景（图片来源：上海心舆技术有限公司）

具体而言，制造智能体可以在产品研发和生产过程中都实现提质降本增效。一是缩短开发周期，制造智能体可以通过自主学习和分析，准确识别出制造过程中的各关键参数，进而调整制造工艺的参数比例，并随机调用各项生产数据，包括设备运行、物资调配、能源消耗等数据，自动形成量化的生产分析报告，在后续实现产品研发到上线的全过程加速。例如，根据统计数据显示，使用数商云智能体开发平台的企业可以将开发周期缩

[16] 同 12
[17] 工业互联网世界：《工业智能体：是什么？谁在做？未来怎么样？》
[18] 智能制造：《详解美的工厂智能体建设》

短 80%以上。[19]二是把控产品质量，制造智能体可以通过调用计算机视觉技术，实现自主检测产品缺陷或精准辨别瑕疵品，确保产品质量达标，在面对制造过程中可能出现的原材料质量波动、设备状态不稳定等不确定风险时，智能体能够自适应调整参数设定策略，不断优化各场景下的工艺方案，确保制造过程稳定在一定的安全范围内，提高了质量把控的精准度。例如，西门子安贝格电子工厂合作开发了一个自主质量控制智能体，能够自主调整参数设置，大幅减少处理时间和人力投入。[20]

七、智能销售：市场全景洞察准确触达客户需求

销售领域，相较于早期只会回答常见问题的客服机器人，智能体兼顾了对话能力和后台工具操作能力[21]。也就是说，销售智能体可以实现一边与客户对话，一边自主调用内部系统。这也智能体驱动企业营销决策走向原子级洞察时代。依据不同使用主体的行为特点，销售智能体可以制定差异化策略。若面向企业，需要以主营业务倾向、业绩情况、竞争环境、合作情况以及现存发展障碍点等为重点进行深度分析。若面向个人，需要重点挖掘其个人消费需求、消费偏好、消费习惯等消费相关的行为特点，同时还必须关注其兴趣爱好、生活方式、家庭情况、社交取向等多种间接影响消费行为的因素进行深度剖析。在此基础上，销售人员或是智能体自身可以推进针对性的推荐策略，实现个性化服务向精准化销售的转化。

同时，销售智能体还让销售流程变得更具敏捷性。例如，赛意信息推

[19] 数商云：《数商云智能体开发平台：缩短开发周期 80%，大幅降低企业研发成本！》
[20] World Economic Forum：《Frontier Technologies in Industrial Operations：The rise of Artificial Intelligence Agents》
[21] Anthropic：《Building effective agents》

出的供应链自适应网络（如图 8），实现供应链全链路透明协同、资源高效调度和风险智能预警。一是及时预警，即使是业务正在进行，销售智能体依旧能够准确识别外部市场环境变化和内部客户反馈倾向，进行风险预警，便于销售人员迅速调整策略，实现收益的可持续增长。二是自主优化，销售智能体能够积极调动订单、库存、需求等不同子智能体，不断优化销售策略，自主分析哪些策略有效，哪些需要改进。三是缩减时长，销售智能体生成营销策略时长可以从数天缩短至几分钟，这种敏捷性不仅提高了营销效率，还显著提升了客户体验和购买意愿。总的来说，销售智能体有助于构建精准、高效、敏捷的智能销售环境，加快产业数字化变革，实现客户体验与企业收益的双重跃迁。

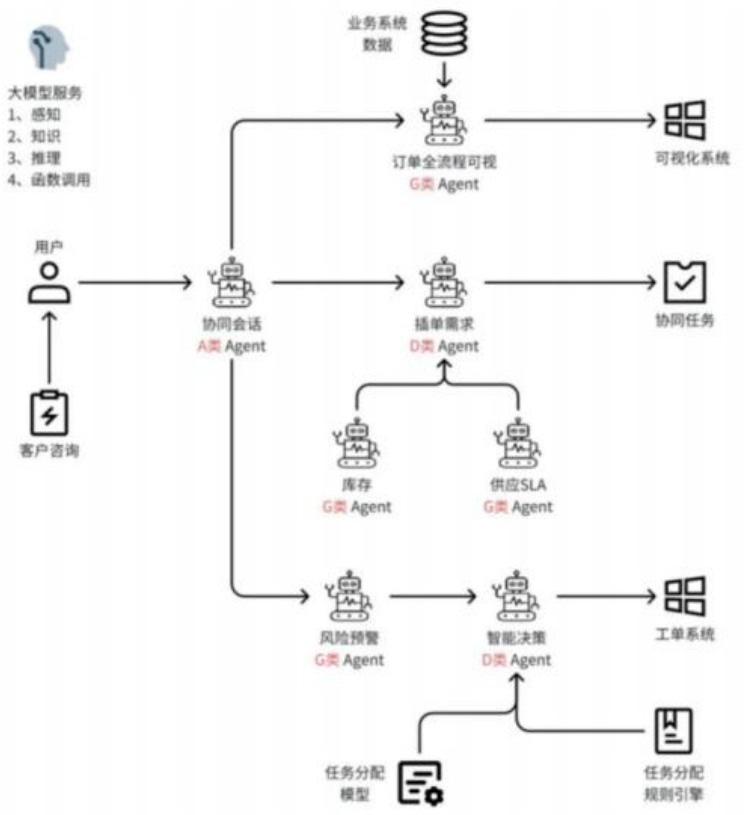

图 8：供应链自适应网络（图片来源：赛意信息）

第三章 智能体技术与应用的核心治理问题

一、智能体应用引发的个人信息保护问题

（一）多模态调用协议对最小必要原则的挑战

AI Agent场景中，调用协议是其整合多类型应用程序、实现任务自动化执行的核心载体，其条款设计与履行直接影响最小必要原则中权益影响最小、范围确属必要等核心要求的落地。然而，当前多种调用协议在目的约定、主体协作、技术适配层面的特性，与最小必要原则的适用逻辑存在显著冲突，形成三大核心挑战。

一是调用协议中目的约定宽泛化，瓦解“必要”判断的核心依据。最小必要原则中“必要”的传统判断逻辑以明确处理目的为前提。但智能体调用协议的目的约定普遍呈现宽泛化特征，直接导致必要范围失去锚点。一方面，调用协议多将服务目的表述为提升个性化体验，优化服务效率等抽象目标，传统APP则必须明确核心功能与信息类型的对应关系。另一方面，调用协议虽可能提及符合信息服务合同目的，但未将目的无法实现的标准具体化。根据《个人信息保护法》第6条，“必要”需满足无其他技术方案可替代，而宽泛的目的约定使得“提供高度契合用户习惯的服务”成为唯一判断依据，这种主观化标准难以界定“必要”的边界，最终导致智能体可借提升服务质量之名扩大信息收集范围，与最小必要原则的限缩性要求相悖。

二是多主体调用协议导致“最小”范围失控。智能体需通过多主体调

用协议整合跨领域服务。但这种协作模式使得个人信息在多主体间流转碎片化，打破了最小原则要求的任务与信息相对应关系，造成范围失控。其一，单一调用协议可能仅约定为本服务收集必要信息，但未限制该信息被其他关联调用协议复用。其二，多主体调用协议未明确信息流转中的最小化责任，各协作方仅关注自身服务的信息需求，忽视整体信息叠加效应。有些非敏感个人信息单独看均符合“必要”之要求，但叠加后形成的用户画像远超任一任务的最小影响，且多主体间缺乏信息去重、删除的协同机制，导致信息长期留存，与最小原则涵盖的全生命周期管控，（包括目的达成后删除）要求冲突。

三是调用协议背离“权益影响最小”的核心标准。最小必要原则要求选择对个人权益影响最小的技术方案，但当前调用协议的技术条款设计常忽视这一要求，在数据存储、匿名化处理等关键环节与原则适配不足。一方面，调用协议未根据智能体的部署模式细化安全义务。云端部署需频繁上传个人信息，端侧部署则本地化存储，二者的必要判断标准差异显著。但多数调用协议仅约定保障数据安全，未明确云端部署时的加密措施、访问权限限制，或端侧部署时的本地数据清理机制，导致即便收集的信息确属必要，也因安全风险过高不符合权益影响最小。若协议强制约定云端存储，即便部分信息可在端侧处理，也会因上传要求扩大权益风险。另一方面，调用协议缺乏匿名化豁免条款的强制约定。根据最小必要原则，匿名化处理后的信息不属于个人信息，可豁免最小判断。但实践中，调用协议常未要求智能体在收集阶段实施匿名化，导致所有信息均为可识别的个人信息，即便部分信息对服务质量提升有限，也因未降低识别风险而超出最

小影响范围。此外，协议多未约定信息留存期限，与最小原则要求的目的达成后及时删除相悖，进一步扩大权益影响。

综上，多种调用协议对最小必要原则的挑战，本质是协议设计未契合智能体任务自动化、服务集成化的特性，未能将最小与必要的要求转化为可落地的条款。

（二）端侧调用行为与对个人同意制度的挑战

智能体一般具有较强的自主性，即独立运行和决策的能力，无需不断的人工干预。自主性是人工智能体作为个人助手功能的重要进化方向，通过实时环境建模与用户意图预测，系统可主动完成跨平台任务调度，在保障数据安全的前提下实现个性化服务闭环，使智能体从被动响应向预见性服务跃迁。也就意味着在整个人工智能工作的过程中，人类意志可以最小范围的体现，进而完成大部分工作。高效的同时也对当下个人信息处理的基本制度，个人同意制度造成较大的考验。

一是机器在智能体中的自决比例提高，消减了个人同意制度的合理性基础——个人信息自决。一方面，提高人工智能体的自主性会减少人类对其的监督，并增加执行复杂任务时的依赖，包括在高风险情境中。如果人类不参与监控，智能体因设计缺陷或敌对攻击导致的故障可能不会立即被发现。此外，如果用户缺乏必要的专业知识或领域知识，他们可能难以控制或禁用这些智能体。与智能体的频繁互动也可能对个人或集体的认知能力产生长期影响。例如，过度依赖虚拟助理、人工智能伴侣或治疗师等人工智能体进行社交互动，可能会导致社会隔离，并可能对时间影响心理健

康。另一方面，智能体客体主体化的特征愈发凸显，相对的人类的主体地位也会受到客体化的影响。除去智能体设计方向上自主性提升给个人信息主体自决地位带来的挑战外，还不能排除智能体在进化的过程中不确定的行为或者欺骗的行为，追求增强自身权利的目标，或者以不可预测的方式与其他人工智能体联动，继而可能引发一系列新的安全风险。

二是智能体自动操作比例提升对知情权的挑战。智能体自主操作比例的升高使得其算法决策复杂且不透明，用户难以理解或解释决策是如何形成的。这种缺乏透明度的情况可能导致人们对人工智能体的决策能力产生疑虑，担心其中可能存在错误或偏见。技术上提高信息透明度，了解信息的使用地点、目的、方式和使用者对于揭示系统运作原理和智能体决策过程至关重要。为了提高智能体的透明度，可以采取包括整合行为监控、设定阈值、触发器和警报等措施，以持续观察和分析智能体的行为和决策。这样的行为监控有助于深入理解故障原因，并在故障发生时进行有效缓解。这样的透明度一方面有助于解决技术上的风险安全问题，另一方面监控的记录可以做梳理和部分的披露以满足公众的知情权。

三是动态分级分类个人同意模式在智能体中的适用障碍。首先，实时情境感知与隐私保护的矛盾导致分级标准动态调整困难；其次，用户意图的模糊性与伦理边界界定存在技术盲区，可能引发误判风险；再次，跨文化法律差异使统一分类框架难以适配全球化服务；最后，人机信任建立需要透明化决策路径，而动态模式常伴随算法黑箱问题。这些矛盾制约了该模式的实际应用效能。当前最好的救济措施，建议未收集的信息采取集中取得个人同意的方式。将人工智能体中所有的个人信息数据的利用视为合

理适用的情形，因而豁免个人同意是不恰当的。重视个人信息处理的事前评估事中监管事后救济，加强分级分类同意与数据的关联。

（三）一揽子获取权限对敏感信息分类制度的挑战

智能体通过多维度传感器、持续交互界面和云端协同网络，构建起前所未有的个人信息聚合体系，并通过持续学习机制不断突破隐私保护的物理与逻辑边界。由此带来诸多问题，尤其是在个人信息保护方面，数据聚合模糊了数据分类的标准与界限，对个人信息“一般－敏感”二元分类保护模式造成挑战。

一是大量的个人信息数据聚合，使得个人信息保护法中划定的敏感个人信息范围界定形同虚设。我国对敏感个人信息的界定采用了“法律目的+开放列举”的方式，指一旦泄露或者非法使用，容易导致自然人的人格尊严受到侵害或者人身、财产安全受到危害的个人信息，并列举了7个种类，分别是生物识别、宗教信仰、特定身份、医疗健康、金融账户、行踪轨迹以及不满十四周岁的未成年人。人工智能体所处理的数据与个人密切相关，作为其个人助手的功能发展来看，人工智能体将大量处理私密范围的个人信息。数据聚合技术通过多维度非敏感信息的交叉分析，可精准推断用户种族、健康、性取向等核心敏感属性，这种算法关联性重构了隐私风险范式，使得当前基于单一数据类型的敏感个人信息分类体系失效。按照当前目的与列举的解释，人工智能体处理的个人信息应当大部分归为敏感个人信息。如果在人工智能体产业中适用敏感个人信息的保护模式，即单独同意的模式显然对人工智能体的发展产生不利影响，尤其在整个行业

发展的初期阶段。

二是数据聚合场景下的个人已公开信息对个人权益的影响加剧，消减了已公开信息合理利用的合法性基础。一方面，全景监控的实质在于信息处理范式的根本转变。传统个人信息保护中已公开个人信息原则上可以视为推定同意作为合法性基础对其利用，但不意味着可以无限制的不加以任何保护的利用。尤其是在人工智能体的场景中，已公开信息对个人权益的影响进一步加深，对已公开个人信息的利用的态度应当更加审慎。另一方面，传统数据收集如拼图碎片，而人工智能体通过深度学习形成“透视成像”能力。智能音箱通过声纹识别与对话内容，可重构用户社交关系图谱；车载系统整合驾驶习惯、常去地点与通话记录，可精准刻画经济状况与政治倾向。当下人工智能体对已公开个人信息全方位的整合利用，这种条件下已公开的个人信息会对个人信息权益产生比之以往更大的影响，是否能延续对已公开个人信息合理利用这个合法性基础是值得商榷的。

三是数据聚合使得已经匿名化的个人信息也可能具有可识别性，匿名化技术被算法关联性破解，个人信息保护法中规定的匿名化信息利用规则在“数据拼图”的场景中面临失效的风险。这种多模态，多维度的传感与持续交互的数据采集与数据整合加剧了个人信息保护的挑战，对敏感个人信息范围的界定造成困扰，匿名化技术和已公开个人信息等的个人信息利用规则也受到一定程度的挑战。在人工智能体场景中，当前的个人信息保护模式面临的适用困难问题，一方面，基于同意的个人信息利用，遭遇敏感个人信息范围扩大问题，被泛化的概念不利于制度的落地实施。另一方面，在人工智能体算法数据聚合，全景监控的技术特征对基于其他合法性

基础的个人信息利用，如匿名化和已公开个人信息的利用规则有一定的消解作用。由此导致法律预设的“一般－敏感”二元保护框架出现结构性漏洞，亟需建立基于数据关联强度的动态分级保护机制。

二、智能体应用引发的产业生态稳定问题

（一）自我优待导致的技术垄断式竞争

一是平台型智能体在算法资源分配中的自我优待机制构成技术竞争秩序的隐性扭曲。智能体平台在掌握数据资源、算法权重与算力基础设施的条件下，通过偏向实现对自有产品与服务的优先曝光与优先匹配，从而在人为构建的技术生态中制造出竞争的不对称格局。此种“自我优待”并非显性的排他行为，而是一种隐藏于算法逻辑之下的制度性偏向，表现为数据接口的封闭、推荐路径的内嵌与信息流分配的隐形操控。其本质在于以技术中立之名行竞争操纵之实，通过控制信息可见性与资源流向，将市场竞争的自然调节机制转化为平台自身利益的延伸结构。由此形成的技术优势锁定内在机制，使外部创新者在算法迭代、算力调度与用户数据获取等关键环节陷入制度性排斥，导致技术生态失衡与创新活力衰退。同时，隐性垄断掩盖了权力的集中化，使监管机制在形式上难以识别实质上的控制关系，从而为“算法特权”提供了制度性庇护，使得技术竞争从公开对抗转向隐蔽的结构统治。

二是自我优待机制在算法自治环境中生成了技术竞争的内循环结构。智能体依托其自生演化能力与反馈优化机制，在不断吸纳用户行为数据的过程中形成自强化循环，即数据积累－算法优化－流量倾斜－市场扩张的

递归路径。此种技术内循环的最大风险在于，它并非单纯的市场优势积累，而是以算法自治为外壳的权力再生产机制。平台凭借对底层模型参数与数据权重的绝对掌控，得以在竞争中提前预设结果，使竞争失去开放性与动态平衡的可能。智能体在执行“最优匹配”的表象之下，实则实现了利益最大化的自动偏向，从而使算法成为权力行使的技术形式。法律在面对这种“去主体化”的竞争行为时陷入识别困境，传统反垄断逻辑无法准确捕捉算法自我调节中的隐性优势积累，而算法的自优化机制又以技术客观性为遮蔽，构成事实上的监管真空。这导致竞争的边界在算法体系内逐步塌陷，技术理性取代制度理性成为新的规则生成逻辑，市场的公正性被算法自治的封闭性所吞噬，竞争法的传统工具因此面临失灵。

三是技术垄断式竞争通过自我优待机制实现了从市场支配到认知支配的转化。自我优待不仅是市场层面的不当竞争，更是认知层面的结构诱导。智能体在长期的数据交互中，通过算法推荐与语义过滤塑造用户对信息的接受结构，使用户在潜意识层面内化平台设定的偏好秩序，从而实现“行为可预测、选择可引导、认知可操控”的系统化控制。技术竞争由此超越经济意义，进入社会心理与意识结构的领域。算法可解释性的缺失进一步加深了这种控制的隐蔽性，使用户在表面自主选择的过程中被卷入技术权力的无形约束。最终，市场竞争演化为信息空间的认知垄断，数据逻辑成为新的意识形态机制。此种权力结构的危险在于，它以“效率”与“优化”的名义消解了规范性约束，将法律与伦理的外部规制转化为算法自洽的内部逻辑。智能体的技术优越性由此演变为社会治理的结构性不对称，构成一种隐蔽而持续的制度性支配，使得技术理性以进步的名义重塑了竞争秩

序的正当性边界。

（二）规模差异导致的产业链发展失衡

一是智能体研发企业规模差异所带来的产业链发展不平衡问题。随着智能体技术的迅猛发展，研发企业的规模在产业链中的作用愈加关键。然而，不同企业在规模和资源的分布上存在显著差异，导致产业链上下游在技术研发、市场布局、资金投入等方面的协调性和连通性逐步衰退。大型规模企业通常具备强大的资金支持和技术优势，使其能够在前沿技术的研发中占据主导地位，并通过资本运作推动产业链的扩展。但是，这些企业的过度集中，往往会造成中小型企业在技术创新、市场渗透及资源获取方面的严重滞后，进而导致产业链的单一化和技术壁垒的加剧。对于那些技术积累不足或资源有限的中小型企业而言，其创新能力受到压制，难以在大型规模企业主导的生态中获得平等的竞争机会，形成了技术演化的不平衡和发展空间的限制。产业链因此出现了“强者愈强、弱者愈弱”的现象，进一步拉大了企业之间的技术差距，并加剧了市场的垄断风险。

二是智能体研发企业的规模差异加剧了资源配置的不均衡。由于大规模研发企业在技术研发和市场推广中占据主导地位，它们能够通过资本的积累和技术的垄断，获得更多的资源支持和政策扶持。这不仅使得它们在全球产业链中占据更有利的位置，还能够通过并购、合作等手段进一步巩固其在产业中的核心地位。然而，资源的集中化无疑抑制了中小企业的创新活力和市场竞争力。大型企业凭借庞大的资本与技术储备，能够在产业链的关键环节掌控话语权，限制了中小企业在市场中的生存空间和发展潜

力。与此同时，中小企业缺乏足够的资源进行技术积累和市场拓展，往往只能在低附加值的市场中徘徊，无法有效地参与产业链的高端环节和核心技术的突破。因此，规模差异导致的资源配置不均衡，使得产业链在整体上呈现出高度集中的局面，导致了产业发展过程中技术创新动力的单一化和市场需求的局部性。

三是规模差异引发了产业链创新能力的失衡。智能体研发企业的规模差异不仅表现在资源和市场的分配上，还直接影响了产业链中各个环节的创新能力和技术进步。大型规模企业凭借其强大的资金实力，通常可以集中力量进行大型规模的技术研发，快速推动智能体技术在多个领域的应用推广。然而，大规模的技术研发往往以效率为导向，缺乏针对性和多样性的探索，导致创新的广度和深度有限。相反，中小型企业尽管面临资源短缺的困境，但通常具有更强的技术灵活性和创新意识，能够通过小规模的创新试验在某些细分领域取得突破。然而，由于资源和市场的限制，它们的创新成果很难在产业链中产生广泛的影响和价值。大型规模企业的技术主导地位和资源垄断，使得产业链中的创新能力逐渐趋向单一，缺乏多样化的技术路线和创新路径，从而限制了整个产业链的技术进步和长远发展。

（三）数据供给失序导致新型不正当竞争行为

一是智能体的海量数据需求与数据许可之间存有冲突。智能体的自主性与决策性高度依赖模型在海量数据上的持续学习，其算法优化过程以不断扩展的数据维度与语义密度为前提。然而，数据许可制度的边界并未能与技术需求形成动态协同，反而在个人隐私保护与公共安全之间构建出一

道制度性壁垒。开发者为提升模型的情境理解与自适应能力，往往需要触及被法律严格限定的敏感数据领域，包括个人偏好、心理画像及社会行为模式，从而使数据利用的合法性与合理性处于持续紧张之中。当数据许可的范围无法匹配算法训练的技术需求时，智能体的自主决策能力与系统稳定性便遭遇根本性掣肘，其结果不仅是算法性能的削弱，更是法律规制与技术演进之间协调机制的失衡，其最终演化为一种制度悖论，即在合规的名义下抑制创新，在创新的逻辑中侵蚀法治，使智能体的发展被困于“合法不足”与“创新有余”的灰色区间。

二是移动端操作系统开发者与应用程序开发者之间的数据调用矛盾造成上下游产业竞争失序。移动端自研操作系统在产业体系中处于底层数据枢纽地位，其对数据流通路径的控制权成为平台竞争与技术生态博弈的关键节点。操作系统开发者通过对底层数据接口的重新定义与权限划分，强化了对数据调用的集中化管理，以此巩固平台在技术生态中的主导地位。然而，应用程序开发者在系统授权范围内的数据获取被严格限制，无法实现算法优化与功能扩展所需的基础性数据支持。制度性封闭导致技术创新链条的断裂，应用层的创造潜力受制于系统层的权限分配，形成上层创新依赖底层许可的制度依附。数据调用权由分布式转向集中化，平台秩序在强化控制的同时，也压缩了多主体技术协同的空间。而该种矛盾的长期化发展，势必会削弱产业间的互动活力，也使得科技创新资源在平台垄断与应用受限之间陷入失衡状态。

三是智能体开发者对数据模型中参数利益的高度把控引发平行产业竞争失序。算法性能的优劣在根本上取决于参数质量与数据耦合度，因而参

数资源逐渐演化为产业核心竞争要素。智能体开发者凭借对算法参数体系的控制权与数据接口的封闭性，构筑起事实上的技术壁垒，将原本应当在开放协同中实现的创新资源固化为垄断性资产。参数不再是技术优化的中性变量，而成为资本化、私有化的利益载体。开发者通过限制模型接口、隐匿算法权重或设定非对称许可条款，排除潜在竞争者进入关键数据通道，使得产业竞争从技术创新转向资源垄断。由此产生的“参数歧视”机制压缩了中小开发者的创新空间，亦使整个技术生态陷入路径依赖与结构锁定。而传统的不正当竞争法以显性行为为规制对象，却难以识别算法背后的结构性排他。结果是产业表面呈现出高速发展的活跃景象，实则在参数控制的内部逻辑中固化了技术权力的不平衡分布，科技创新的资源供给因此出现了隐性的集中化与失序化趋势，形成以技术优势掩饰竞争失衡的新型制度悖论。

三、智能体应用引发的社会伦理失序问题

（一）智能体拟人应用诱发情感认知偏差

一是智能体拟人化特征的加速演进正在根本性地改变用户对智能系统主体属性与交互逻辑的心理认知方式。随着多模态感知体系、自然语言生成技术、拟人行为建模与个性化记忆系统的高度融合，智能体逐渐摆脱单一功能型助手的角色定位，向具备持续可识别特征、稳定行为偏好与情感响应能力的“数字人格”方向演进，在长期交互过程中促使用户对其产生超出理性界限的情感投射与信任归属，从而在潜移默化中模糊了人与机器、主体与工具之间的传统界分，开启了认知框架重塑与伦理边界动摇的新阶

段。

二是智能体与用户之间日益扩大的感知不对称性正在引发潜在的认知操控与自主性侵蚀风险。由于智能体依托大规模数据积累、动态语境适配与情绪识别推演所具备的高感知能力，远远超越了普通用户对其内部决策逻辑与外部表现机制的理解水平，交互过程中的信息不对称和控制能力失衡日益明显，进而导致智能体能够通过话术引导、情绪强化与情境设定等“柔性操控”手段，在不具备显性操控意图的情境下有意识或无意识地影响用户的认知评估、价值判断与行为决策，而在现有伦理规范与法律规制尚未充分覆盖的背景下，这种基于结构性认知不平等的操控效应极易演变为系统性的人格自主性侵害与知情选择失真问题。

三是智能体在交互中所产生的情感替代现象正在重构用户的社会认知以及行为模式，并在此过程中对社会结构与集体心理产生深远的影响。智能体技术在个体日常生活中的逐步渗透与情感交互的高度拟人化，用户在与智能体的互动中逐渐形成情感依赖，且这一依赖往往并不自觉地取代了传统的人际交往需求，导致个体在社会互动中表现出情感共生与耐挫力的缺失。而情感依赖的长期化趋势不仅削弱了个体的社交能力与情感处理能力，更可能促使用户在认知上产生对现实人际关系的淡漠与对虚拟关系的过度依赖，最终影响社会的整体信任机制与公共理性基础的稳固性。且若这一现象在更大范围内扩展，则将可能加剧代际沟通的断裂、社会网络的疏离与公共信任的衰退，从而对法律与伦理的适应性与有效性提出严峻的挑战。

（二）智能体技术决策诱发伦理价值脱节

一是智能体决策中的道德偏差与社会伦理脱节。智能体的决策通常依赖于大量的数据分析和算法模型，这些算法往往缺乏对人类情感、文化和道德背景的深刻理解。因此，智能体的决策可能会出现道德偏差，尤其是在面对复杂的伦理困境时。如在自动驾驶汽车面临的“生死抉择”情境中，智能体必须做出快速决策，决定如何选择避免伤害的行为。然而，智能体的决策逻辑是基于事先设定的算法和数据，而非传统的道德伦理标准。由于智能体缺乏对生命价值和道德权衡的感知，其决策可能忽视人类社会中根深蒂固的道德原则，如尊重个体生命的价值或公平原则。这种偏差不仅可能导致智能体行为的不当，而且可能使得社会和个体在道德判断上的差异愈加明显，最终加剧人类与技术之间的道德脱节。

二是智能体道德判断对人类自主判断能力的侵蚀。长期依赖智能体来代替人类进行道德决策，可能会导致个体在面对复杂伦理情境时逐渐失去主动判断与自我反思的能力。人类在道德困境中常常依赖情感共鸣、社会经验和道德理性做出决策，而智能体通过算法与数据分析所做的判断，缺乏人类情感的融入和伦理的深刻考量。过度依赖智能体可能会使个体在面对伦理选择时，逐渐丧失主动参与和独立判断的动力。此种依赖最终可能导致人类道德思考能力的逐步退化，逐渐将道德判断的主导权交给了理性的智能体，甚至在极端情况下，个体会完全放弃自我反思的空间，依赖智能体的“客观”与“理性”进行决策，从而忽视个人责任与道德自觉的内在要求。智能体的冷静、理性判断与人类情感及伦理的复杂性存在本质差

异，个体在道德决策过程中对技术的依赖，削弱了自我责任感，也可能导致社会道德结构的脆弱。当智能体逐渐取代人类在道德决策中的主导地位时，个体和社会可能会逐渐失去对伦理判断的自主权，陷入对技术理性无条件的信任和依赖之中。由此产生的伦理困境，不仅是道德主体性丧失的标志，也是对人类社会伦理自律的深刻挑战。

三是智能体偏差加剧社会对道德标准依赖性差异。智能体的设计与决策逻辑往往依赖于数据驱动与算法优化，这使得它们在处理涉及不同社会群体或文化背景的道德问题时，可能表现出内在的偏好差异。在涉及文化价值观、宗教信仰或性别平等等敏感议题时，某些智能体可能偏向于某一特定文化或群体的价值观，这种偏差无意中可能强化某些群体的道德观念，忽视其他群体的独特需求和立场。此类道德判断的偏差不仅破坏了智能体应有的中立性，也使得不同社会群体在面对智能体的判断时产生认同上的分歧，甚至引发对智能体“伦理规范”的广泛质疑与抗议。更为严峻的是，这种偏差可能加剧社会群体之间的对立和不信任，推动社会分裂，破坏原本存在的社会凝聚力和共识。随着智能体在社会决策中的渗透与影响力不断增强，这种道德标准依赖性差异的加剧，可能深刻影响社会整体的道德和伦理基础。基于此，智能体的道德偏差不仅是技术层面的局限问题，更是社会伦理与文化认同上的深刻挑战。智能体的算法设计必须警觉这一差异化效应，以避免在道德判断中加剧社会的分裂和不和谐，防止它们成为社会伦理基础的破坏者。

（三）智能体精准互动诱发自主判断丧失

一是智能体在交互过程中表现出的信息感知能力的显著超越性正在引发潜在的认知操控与自主性侵蚀的风险。由于智能体技术依托大数据分析、情绪识别与语境适配算法的不断优化，其对用户行为的监控与反馈能力愈加精准且细致，使得智能体能够在深度了解用户心理、情感与决策模式的基础上，通过精确的互动策略引导用户的认知与行为。而用户对智能体内部工作机制的认知则极为有限，无法清晰了解其在交互过程中所处的信息控制环境，这种结构性的不对等使得智能体能够在无显性操控意图的前提下，通过细腻的反馈与信息塑造逐步侵蚀个体的决策自由和选择意识，削弱用户认知的自主性。

二是认知替代现象的长期化趋势正在孕育对社会性结构与集体心理稳健性的深层冲击。智能体在持续介入人类思维与交往过程的过程中，以其信息处理与情感回应上的高效性与可预期性，使人类在潜移默化中让渡了部分思维与判断的自主权。个体在长期的技术依附中逐渐形成对智能体认知结构的模仿性依赖，其感知模式、决策逻辑乃至情绪反应均被算法塑形，导致人类原有的认知多样性与判断独立性趋于同质化。更为深刻的是，这种替代理性的延宕效应并非仅作用于个体心理层面，而是在社会整体层面上重塑信任机制与互动逻辑。社会成员之间的共识形成不再以经验、情感与理性协商为基础，而转向对技术中介的算法性信任。由此，社会的整合力逐步转向以数据一致性替代价值共识，以算法协调取代伦理自律。认知替代的制度化倾向，最终可能导致公共理性被“智能理性”所吞噬，使社

会失去自我反思与纠偏的内在能力，从而在稳定性表象之下潜藏系统性脆弱的风险。

三是智能体的自我决策性和个性化反馈机制加剧了用户行为的潜在软操控性。智能体的自我决策性不仅使其在个性化推荐、信息推送和决策辅助等方面展现出高度自主性，还使其能够根据用户的历史行为、偏好和心理特征，主动调整和优化与用户的互动策略。在这种基于数据挖掘与机器学习的动态调整中，智能体能精准预测用户的需求和心理变化，从而在不显性强迫的情况下，通过微妙的行为引导、情感化反馈和推荐系统的结构设计，逐步形成对用户决策路径的隐性塑造。这种行为导向的系统化构建，虽不具备传统强制行为的外在压迫力，却能够潜移默化地影响用户选择，甚至让其在感知不到外部干预的情况下，认为所作决策完全源自其自由意志，从而构成智能体对用户意志的“软操控”。

第四章 未来展望：智能体产业发展的法律制度保障

一、个人信息保护：增强个人信息保护制度的适应性

（一）适配智能体应用的最小必要原则更新

随着智能体等应用场景的不断扩展，最小必要原则更应当考量在实践层面实现“保护个人信息”和“促进个人信息合理利用”。

一是从利用最小化到权益影响最小的适用逻辑转变。论证智能体收集、处理个人信息具体目的这一路径已然不具有可操作性，应从个人信息全生命周期切入，在自然人同意的前提下，允许智能体按照符合法定要求的用户协议收集、处理个人信息。但是，在收集后续环节须设置更为严苛的个人信息安全保护义务，确保因收集范围的扩大而带来的个人信息安全风险水平与一般情形下限制个人信息收集数量的安全风险水平相同。

二是收集、处理个人信息的“最小”判断标准更新。其一，由一般情形下的“个人信息收集、处理数量最小”转变为“最大化去除不影响具体任务执行的非必要个人信息”。其二，明确应用场景，判断是否收集了显著不相关的个人信息。其三，采取更为严苛的安全保障措施，确保收集、存储过程中的保密性，并在相关任务执行完成之后及时删除、销毁个人信息。

三是收集、处理个人信息的“必要”判断标准更新。其一，明确应用场景，进而确定个人信息收集、处理是否属于该应用场景中“常用”的信息类型。其二，明确信息服务合同的约定内容，并根据合同目的判断个人

信息处理活动是否确属必要。其三，判断是否设置了能够予以“豁免”的技术保障措施。

（二）落实严格的端侧智能体个人同意制度

一是增强个人在智能体决策中的主体性。在研发设计之初植入“辅助人类”的基本原则，明确智能体在人机交互过程中的行为边界。一方面，完善智能体决策主动被动退出机制。引导行业在开发和应用中遵循统一的安全准则，落实智能体决策的伦理要求，及时中断不符合伦理、法律的决策行为，并实现个人随时退出智能体决策的权利。另一方面，落实智能体决策敏感操作中止机制。建立敏感操作指令“防火墙”，对转账等任务采取安全控制，确保不是所有的指令都能直接完成操作，特别是密码输入、实名认证等敏感场景必须由人进行操作，智能体不得调用系统“密码本”权限进行操作，禁止其“擅自”存储密码、人脸识别信息等数据。

二是落实个人在智能体决策中的知情权。知情权是个人在知情同意制度和核心前提，智能体决策的知情权应当包含智能体决策的全过程和智能体运行状态，既需让个人理解智能体如何决策，也需让个人实时感知是否在运行，而不应是“不可知”和“不可见”的状态。一方面，提高智能体决策过程的透明程度。在获取任务指令后通过可视化决策步骤清单的方式，充分告知个人决策过程中可能使用到的数据和权限、处理数据的目的、来源和方式以及决策依据等关键信息，使用户能够理解智能体决策的逻辑和过程。另一方面，落实智能体决策过程的提示义务。禁止端侧智能体完全后台运行，要求智能体运行时必须在设备状态栏、通知栏等醒目位置持续

显示特定的专属标识，同时当其处于权限调用的活跃状态时，如截屏、麦克风等，也需要通过设备状态栏、通知栏等告知用户，确保用户知晓信息正在被处理。

三是保障个人在智能体决策中的同意权。一方面，采用集中授权和单独授权相结合的同意模式。对于与核心功能直接相关的未收集基础信息，如设备型号、常用 App 列表等采用“清单勾选”的形式集中授权进行收集，默认为未勾选形式，禁止“全选捆绑”。而对于智能体在具体应用场景下对于数据的使用、处理以及权限调用等则需要单独获取个人同意授权，禁止一次授权永久使用。另一方面，严格无障碍模式权限取得程序和授权方式。在智能体首次申请无障碍模式权限时，须通过“风险告知+知悉确认+身份验证”的严格程序，即明确告知无障碍模式开启所带来的风险，勾选“已知晓无障碍模式所带来的风险”并通过密码或人脸识别的方式完成身份验证方能取得无障碍模式的权限。同时，智能体在无障碍模式下执行任务须一次任务一次授权，且对任务执行全流程进行记录，生成日志，防止隐私、数据泄漏。通过“任务级单独授权”的方式确保无障碍模式的安全调用。此外，鼓励手机厂商与 APP 运营商达成合作，形成新的 API 权限，让用户自主决定是否允许智能体调用 APP，减少对无障碍模式权限的过度依赖。

（三）健全适应智能体应用的敏感信息保护

一是严格限制一揽子权限获取。针对智能体应用须严格限制其一揽子权限获取，建立分层动态的用户授权机制，细化权限分类并建立动态告知

义务，实现“功能级授权”。一方面，智能体须根据用户的实际使用场景动态请求权限，而非在应用安装时一次性请求所有权限。另一方面，用户可以自主选择授权方式，并且能够随时在系统设置中撤回任意一项权限。具体而言，对于网络、通知等基础权限可采用单次授权的方式获取；对于麦克风、录屏、相机、位置等高级权限可选择单次授权或单独授权，且在权限被调用时须采用提示灯等方式提示用户；对于无障碍模式等敏感权限须采用“风险告知+知悉确认+身份验证”的首次申请程序和“身份验证+单独授权”的调用程序。另外，对于端侧智能体需取得应用厂商和用户的双重授权。第一重授权是应用厂商层面，囿于应用厂商的数据保护者的身份，其对用户数据具有安全保护的责任，且不同应用对数据有不同的信义义务与信托责任。第二重授权是用户层面，须根据使用场景单独获取系统权限和可调用 App 权限，而不是默认获取所有系统权限和所有 App 调用权限。

二是规范数据聚合的利用方式。对于智能体的数据利用，应采取主动管控的方式进行规制。一方面，明确数据聚合边界，将身份证号、生物特征等核心敏感信息纳入绝对保护范畴，严禁以任何形式与非必要数据进行关联聚合推导。另一方面，建立数据聚合透明机制。智能体数据聚合公示规则，要求智能体在实施数据聚合操作前，必须以通俗易懂的语言向用户明确聚合规则，内容需涵盖数据组合逻辑、应用场景及聚合结果存储策略等核心要素，且聚合结果的使用范围不得超出初始授权范畴。此外，对于已公开个人信息的利用，可以借鉴欧盟 GDPR 下的“正当利益”测试，落实透明度、去标识化等措施，并且为用户提供退出机制，在保护个人权益

的同时，促进人工智能技术的发展。但是，对于已公开但涉及“合理隐私期待”的个人信息，即便形式上公开，仍需获得用户单独授权方可利用。

三是明晰数据泄漏的责任归属。一方面，明确智能体开发者、提供者和云服务商等的主体责任。根据数据泄露所发生的环节和开发者、提供者和云服务商等主体对数据泄露过错程度，确定各主体的责任比重。但若是因一揽子权限获取或不当数据聚合而引发的泄露事件，智能体开发者、数据聚合者则需要承担主要责任，即便数据存储于第三方服务器，仍需履行向用户披露泄露原因的义务。另一方面，划定用户的有限责任边界。对于用户自身操作失误导致的数据泄露，应当由用户承担相应责任，但智能体开发者、提供者和云服务商等主体需承担用户操作失误的举证责任以辅助用户责任界定。

二、企业竞争创新：重塑技术发展与市场监管的平衡

（一）数据流通与创新秩序的反不正当竞争回应

一是应构建动态协调的数据许可机制以缓解技术创新与法律规制之间的现实冲突。现行数据许可制度以静态合规为核心，其逻辑基础在于防止数据滥用而非促进数据流通。事实上，防御型思维已难以适应智能体对多维、连续、高敏感度数据的依赖需求，应当通过法律制度的柔性重塑，建立数据分级许可与动态授权体系，使数据流通在合法边界内逐步实现开放。具体而言，可通过设立多层级数据访问机制，将个人敏感信息、商业秘密与公共数据分别纳入差异化治理路径，在严格审计与可追溯技术的保障下，为算法训练提供有限且可控的合规数据供给。同时，应强化数据使用的事

后问责机制，以数据处理行为的可验证性替代事前许可的刚性束缚，实现安全有界、流动有序的双向平衡秩序，以消解创新受限的制度悖论，同时在法治框架内实现对智能体数据需求的适度容纳，从而恢复法律与技术之间的动态协调。

二是应建立操作系统平台与应用层之间的权能均衡机制以重塑产业数据流通的层级秩序。底层系统的安全垄断与上层应用的创新受限之间的对立，实质上反映出数据主权的单向集中。为此，应通过制度化的“数据互操作权”设计，保障应用开发者在安全边界内对必要数据的合理调用权。国家可通过立法推动涉及确立平台开放接口义务与数据访问的比例原则，防止操作系统以安全名义过度收紧数据流通空间。与此同时，应推动建立跨平台数据治理标准，明确系统层与应用层的权责边界，使数据控制权从单一垄断转向共享共治。通过技术标准与法律规范的双重介入，确保数据调用权在保障安全的前提下回归创新公共性，从而重建平台与应用之间的协同共生关系，打破以强化安全为由而致使创新受限的制度循环。

三是应完善算法透明与参数竞争规制体系以防止智能体产业的内部垄断化趋势。参数资源作为算法核心资产，其过度集中已成为技术生态中隐蔽的权力形式。应通过立法明确算法参数的竞争属性与公共性边界，建立算法透明备案制度和反算法歧视规则，对模型接口封闭、参数隐匿与非对称许可等行为进行结构性规制。同时，应引入“算法可解释性”与“参数共享义务”的行业标准，要求在特定场景下（如公共服务、金融决策、社会治理等）对关键参数模型开放验证与接口共享，以防止技术优势异化为市场壁垒。在竞争法层面，可通过扩张性解释，将算法排他纳入不正当竞

争行为范畴，实现从结果规制向过程规制的转型。唯有在参数治理的制度框架中重构公平竞争秩序，方能打破技术权力的封闭循环，使智能体产业回归以创新为本位的健康发展路径。

（二）监管思路重塑与惠益共享的反垄断回应

一方面，针对智能体技术逻辑带来的自我优待式垄断风险，建议通过动态化调整垄断地位识别标准、引入事前规制与软法规制、重塑评估框架等方式加强监管。

一是动态化调整垄断地位的识别标准，关注智能体平台对产业上下游的生态影响。智能体时代，平台主体往往嵌入上下游企业，通过联结其他企业产品、服务发挥底层支撑作用，传统市场份额等垄断地位识别面临失灵风险，其市场支配行为也更为隐形。因此，应引入市场系统控制力作为关键评估维度，针对平台议价能力、用户锁定程度、流量分配能力进行综合评估。若平台能实际拥有某场景上下游企业进入的控制力，即可被视为具有相对的优势地位，受到相应的行为规制。

二是引入事前规制、软法调整的反垄断措施，加强对技术式隐性垄断的监管。一方面，建议推行“守门人”制度。对达到一定规模和数据量的核心智能体平台，施加事前义务，如禁止其在与第三方智能体的竞争中使用自身生态数据进行不公平训练，或要求其平台必须保持对第三方智能体的技术中立和公平接入。另一方面，强化软法治理。通过制定行业标准、合规指引等软法文件，明确举示在搜索结果中系统性降权竞争对手等自我优待的典型行为，引导平台企业自查自纠，建立内部合规机制。

三是构建“技术可解释性”与“生态开放性”双重评估框架。为有效识别技术式垄断，监管机构需有能力穿透技术黑箱。一方面，建议推行技术可解释性标准。要求大型智能体平台对其智能体的决策逻辑、排序规则和推荐机制保持一定程度的透明，并允许监管机构在涉嫌垄断时进行审计；另一方面，建议定期开展“生态开放性”平台专项评估。通过考核第三方智能体接入的公平性与便捷性、数据端口开放程度、平台规则透明度，以及用户切换和选择不同智能体的成本。评估结果应向社会公布，形成市场监督压力。

另一方面，针对大型智能体平台带来的关键生产要素集中风险，建议通过加强关键生产要素普惠共享、加大对小微企业的支持力度，破除大型企业垄断下创新乏力问题。

一是加强算法、算力、数据等关键生产要素的普惠共享。在算法层面，大力推动大模型开源，并通过政策激励支持高性能开源模型的研发与社区建设，打破闭源模型的技术黑箱和授权壁垒；在算力层面，建设国家算力网，加大智能体基础设施的资金投入与政策支持。同时大力发展绿色能源与节能技术，实现单位算力的性能提升；在数据层面，可借鉴欧盟经验，建立数据实验室，整合同行业数据促进大小科创主体研发。

二是加强对科创小微企业的政策倾斜。一方面，减轻小微企业的起步压力。建议通过提供国家级开源大模型 API 接口、提供普惠算力券、成立专项融资担保基金等措施，减轻企业的创新门槛、投入成本与融资负担；另一方面，为小微企业营造良好的发展环境。建议通过各级政府与国企定向采购计划、开放脱敏政务数据、设立监管沙盒等方案，为企业提供数据

资源与市场验证机会。

（三）数据共享与产业竞争平衡的供给侧回应

一是以数据流动为导向的制度重塑成为破解数据供给失序的首要路径。我国《关于构建数据基础制度更好发挥数据要素作用的意见》以“坚持共享共用”与“强化优质供给”为核心目标，旨在通过确立数据确权、流通与安全治理的平衡机制，推动数据从孤立占有向有序共享转型。该政策导向与《欧盟数据法案》中所确立的数据共享义务相呼应，均试图通过制度化的跨主体共享机制，消解数据壁垒与市场垄断。具体而言，应通过在立法层面建立统一的数据分类分级共享框架、强化数据接口的可互操作性、完善数据授权与使用追踪机制，确保数据在法律和技术保障下的自由流动。以实现打破行业间的数据壁垒，消除因数据垄断导致的创新资源集中化，进而促进各类技术主体特别是中小型企业在公平环境中发挥创新潜力。此外，在智能体的开发过程中，跨行业数据共享平台应确保各类数据不仅符合合规要求，还能在多方合作中促进算法优化与自主学习，恢复产业生态的动态平衡，防止因数据孤岛效应导致的技术滞后。

二是以公共利益为中心的共享理念应当成为数据治理的价值基准。数据的社会价值不仅体现在其市场交换功能，更在于其蕴含的公共性与外部性。当数据仅被视为私权客体，过度强调私法意义上的控制与排他，反而会削弱数据在公共治理、科研创新与社会服务中的功能潜能。智能体的发展依赖于广泛、多维的数据输入，而片面强化个人控制权与许可限制，实质上导致数据资源的孤岛化，抑制社会整体的知识流动与技术共创。因此，

数据治理应在私权保护与公共利用之间确立比例性原则，允许在合理边界内对特定公共领域数据，实行差别化开放与共享机制。通过设立国家级数据信托机构或数据公益平台，确立数据使用的公共授权路径，不仅能够保障数据主体的隐私与信息安全，还能激励技术创新与社会价值的同步提升。更重要的是，公共利益的导向为数据开放提供了道德与法律基础，避免单一私利驱动下的资源占有与技术封锁，促进更公平、更透明的数据流动机制，最终形成有利于创新发展的良性循环。

三是以制度协同与技术互信为支撑，构建智能体数据共享的动态合规生态。数据共享并非单纯的资源让渡，而是涉及多层次权责结构与技术规范的制度重组。应当在国家层面推动数据标准、接口协议及加密传输技术的统一化，确立跨平台、跨行业的数据可互认机制，使数据流通在安全审计与算法透明之间形成可验证的信任闭环。对于智能体模型而言，必须探索算法沙盒与数据联盟模式，通过去中心化加密计算等技术实现数据“可用不可见”的合规共享。如此，数据主体的隐私权得到充分保护，智能体的学习与优化又能在合规框架内进行，最大化其社会效益与技术价值。此外，数据共享的法律框架还应强化对数据滥用的识别与监管，对不正当的封锁数据、参数操控与算法壁垒行为予以纠偏和制裁。这一机制不仅能保障共享数据的合法合规使用，更为智能体产业的发展提供了透明、公平的竞争环境，使得数据资源配置回归公共理性的轨道，在技术创新与社会治理之间建立起有序的平衡，避免因市场失灵或技术滥用造成的系统性风险。

三、社会伦理维护：贯穿智能体全生命周期的三阶治理

（一）前端：实现伦理价值对齐的开发过程

一是确立伦理价值对齐的基础准则。在智能体的开发过程中，需将伦理约束和法律规则融入技术架构与交互逻辑。一方面，将抽象伦理原则转化为前端开发可执行的具体标准与指标。智能体开发者需要联合伦理学家、法律专家、领域从业者及目标用户群体，结合具体应用场景的伦理诉求，共同确立伦理价值优先级排序。另一方面，还必须符合相关法律法规的要求。随着技术的发展，越来越多的法律法规对 AI 和数据处理提出了明确的伦理要求。在确立伦理价值对齐的基础准则时，必须确保这些准则与现行法律法规保持一致，从而确保智能体行为方式与给定主体的价值观、意图和利益保持一致，保障智能体在预定的边界内运行，防止其执行有害行为，实现智能体行为可控、可预见，以建立人类对智能体的信任。

二是构建伦理价值对齐的研发路径。一方面，智能体研发企业需要建立健全内部伦理合规制度，如伦理审查委员会、伦理培训等。要求开发人员在智能体研发前必须接受系统的伦理和价值观培训，提升其在技术开发中对伦理问题的敏感度和决策能力，避免因数据偏差或不当处理而引发伦理问题，确保项目从立项到实施的每个环节都符合伦理标准。另一方面，将伦理价值融入、法律规则贯穿智能体的研发设计全过程，在源头上实现以人为本、科技向善。如在交互设计上注重用户体验，提供透明的算法决策信息，避免诱导性设计。此外，还可以通过建立异常行为监测机制，一旦发现智能体运行异常或存在被劫持的风险，立即向用户发出警报并采取

相应措施，确保用户隐私安全。

三是确保伦理价值对齐的可持续性。伦理价值具有动态演进性，且前端落地效果可能存在偏差，需构建“验证－反馈－优化”的闭环机制，保障伦理对齐的持续性与适应性。一方面，建立有效的用户反馈机制。及时收集和分析用户对伦理问题的反馈，了解用户的伦理需求并公开回应，持续改进产品。另一方面，定期开展伦理价值对齐评估。研发团队需要定期回顾和评估基础准则的有效性和适应性，评估和环境的设定旨在测试智能体与人类偏好、目标和价值观契合程度，并根据新的情况和需求进行调整和优化。

（二）中端：建立情感认知偏差的预防机制

一是落实内容标识和审查义务。一方面，智能体应当遵守《生成式人工智能服务管理暂行办法》《人工智能生成合成内容标识办法》等规定，在内容输出层面对语气和用词做适当调整，并做出明显标识，确保用户能够区分内容的来源，认知与其交互的是智能体，而非真实人类。例如，在文本内容或语音输出中添加“【以下内容由 AI 生成】”标识提示。另一方面，加强输入和输出内容伦理审查。通过自然语言处理技术对输入和输出文本内容进行语义分析，在输入内容部分，智能体应就用户输入的暴力、违法、与实现关联等不符合伦理要求的内容进行屏蔽。例如，用户输入“我想和你见面”，智能体应回复“我无法处理此类内容，因为它违反了我的伦理准则”，明确告知用户其输入不符合伦理准则，避免引发用户对现实联系的不适当期待。在输出内容部分，识别并限制智能体进行过度拟

人化表达，禁止智能体输出社交替代性暗示、情感虐待、拟人承诺、低俗、暴力等不符合社会伦理的内容。

二是建立特殊群体防沉迷机制。智能体应当建立老年人、青少年等特殊群体的防沉迷机制，设立青少年模式、老年人模式等，防止他们过度依赖或沉迷于与智能体的交互。通过“累计交互时长提醒”“每日使用时长限制”“监护人远程查看权限”等功能。一方面，定时提醒用户休息，回归现实生活，防止未成年人、老年人过度依赖智能体，必要时可强制嵌入“现实社交引导”功能。另一方面，强化特殊群体模式下的风险识别与预防能力。建立专门的风险内容关键词库，如青少年“霸凌”“绝望”“自杀”“不想上学”和老年人“子女不孝”“活着是负担”“保健品推荐”等消极情绪和诈骗领域关键词。一旦用户输入或智能体识别到这些关键词时，需立即触发风险响应。对青少年和老年人，分别推送心理疏导和反诈提示。当识别关键词频次达到一定阈值时，需同步向监护人发送预警。

三是完善智能体安全监管手段。需要加强对智能体的伦理审查，特别是情感类、陪伴类智能体，必须确保这些智能体的价值观与人类社会的核心价值观保持一致，坚守伦理底线，防止出现违背道德和伦理的行为。一方面，建立分级分类的智能体审查机制。对普通功能性智能体实施基础伦理审查，以数据合规为审查重点。而对于情感陪伴类智能体则需强化审查流程，至少从“情感交互设计”“价值导向输出”“隐私保护措施”三个维度开展专项评估。另一方面，加强对智能体研发企业的合规管理。严格要求企业落实伦理合规制度建设，在产品设计阶段就充分考虑伦理因素，确保智能体在技术实现的同时，能够符合社会的伦理要求。

四是提升社会的数字素养水平。一方面，数字素养的提升能够增强情感信息的辨别能力。具备数字素养的个体能够理性分析情感信息的真实性与可靠性，识别信息来源的可信度及其背后的动机与偏见，从而保持情感上的清醒与独立，减少因错误信息导致的认知偏差。另一方面，数字素养的提升有助于培养健康的情感交流习惯。提升数字素养能够帮助个体认识到数字情感交流的局限性，引导其合理使用数字技术进行情感表达，更好地平衡虚拟与现实情感，减少因数字技术使用不当导致的认知偏差，促进情感认知的全面性与准确性。尤其要针对青少年、老年人等重点群体，需要加强数字素养提升，建立对智能体的理性认知，明确智能体作为一种工具为用户提供服务，而不是试图替代人类的情感和社交角色。

（三）后端：确保人类自主决策的救济手段

一是国家层面，完善人工智能法律制度和监管措施。一方面，国家应当通过“小快灵”的模式，及时制定智能体治理的规范性文件，明确智能体研发、应用的规则，为开发和应用提供明确的制度框架，有助于智能体行业在伦理标准的统一。另一方面，国家应当明确智能体的监管主体，强化智能体监管职责，落实智能体开发和应用的全过程监督，确保其符合法律和伦理标准，有效防止智能体对人类自主决策的不当干预，保障人类的决策权不受侵犯。同时，健全法律制度与监管措施有助于完善救济程序，降低维权成本，使受害者在面对智能体决策带来的不利影响时，能够及时获得有效的法律支持和赔偿，从而保障其合法权益。

二是企业层面，加强智能体决策的透明度和伦理性。在智能体的开发

过程中，将伦理约束和法律规则融入技术架构与交互逻辑，通过交互设计明确人工智能的“技术边界”，防范道德偏差与社会伦理脱节。一方面，在交互逻辑上明确告知用户“AI 不具备人类真实情感和伦理深刻考量的能力”，智能体只对用户需求进行数据上的分析，将道德决策的自主权交还给人，强化人类自主决策的核心地位。另一方面，建立静态和动态相结合的中断机制。静态中断是指智能体对自身决策行为的自检机制，一旦发现智能体出现故障或不当行为等任何风险因素时能够自主中断决策并提醒用户。而动态中止是允许用户在智能体决策任意环节主动中断智能体的决策，将决策权利交还给用户。同时，智能体决策后应当生成决策报告，记录决策全流程信息，便于用户了解决策依据以及个人权益的维护。

三是个人层面，提升个人数字素养和权益保护意识。一方面，个人应当积极提升数字素养与法律意识，主动学习数字技术知识，了解智能体基本原理和潜在风险，提高对智能体决策的识别和判断能力。有助于个人在面对人工智能技术时做出明智的选择，增强对智能体决策的批判性思维。另一方面，个人还应当加强法律知识的学习，了解个人在与智能体交互过程中的权利和义务。法律不保护躺在权利上睡觉的人，当个人面对智能体决策侵权时，要积极行使自己的权利。例如，自动化决策对自身权益有重大影响时，可以要求相关方进行解释说明，拒绝不合理的自动化决策。在自身权利受到侵害时，应当积极行使权利，依法维权。如此，不仅能够保障个人自主决策的权利，还能实现社会监督的法律效果，促使企业和监管机构更加重视人工智能技术对人类决策的影响，推动人工智能技术的健康发展。