安全为基、创新为翼：

构建智能体产业发展新契约

传播内容认知全国重点实验室

2026 年 9 月

目录

前言 1

第一章 人本：锚定用户长期价值、守住人的主体性 3

一、发展特征：个人智能中枢的形成与人机交互的范式重构 3

二、关键挑战：能力放大背后的主体性隐忧 6

三、实现路径：以制度设计与价值重塑守住“人本”底线 8

第二章 安全：筑牢 AI 产业的发展根基 11

一、发展特征：新型安全范式形成 11

二、关键挑战：三级传导下的风险涌现 13

三、实现路径：全生命周期管理下的分级授权与多重防护 16

第三章 透明：可审计、可解释、可追溯的治理基座 19

一、发展特征：全链路可审计成为信任基础 19

二、关键挑战：透明度赤字与信任危机 20

三、实现路径：构建动态平衡的系统工程 22

第四章 普惠：弥合智能鸿沟，守护生态公平 25

一、发展特征：智能体降低使用门槛，新创业范式涌现 25

二、关键挑战：生态公平与产业可持续性 27

三、实现路径：构建智能体时代的普惠框架 29

第五章 协作：完善授权与开放的生态基座 32

一、发展特征：从单体智能迈向多主体协同 32

二、关键挑战：制度缺口与生态失序 33

三、实现路径：构建开放、高效、可持续的 AI 产业协作新生态 34

第六章 创新：安全与发展并重的负责任创新 36

一、发展特征：AI 正改变创新的基本逻辑 36

二、关键挑战：创新不能以透支信任为代价 39

三、实现路径：构建可持续的创新环境 41

第七章 结语：构建智能体发展的新契约 47

前言

2026 年以来，智能体赛道爆款不断，从“龙虾”（OpenClaw）火爆出圈，上线两个月 GitHub 星标突破 30 万，登顶历史第一；到驾驭工程（Harness Engineering）理念的横空出世，为狂奔的 AI 装上了缰绳与护栏。这些现象级事件标志着智能体赛道正在多个方向突破，并带动着智能体向社会和产业领域不断普及。

龙虾的火爆出圈在社会上掀起一股“养虾”热潮。国内人工智能企业纷纷跟进，上线一键部署服务或发布基于自身模型的智能代理类工具。在 PC 端掀起“龙虾风云”的同时，基于移动端的智能体发展同样进展迅速。苹果、谷歌相继发布端侧智能体，国内手机厂商也在智能体上不断探索，一场人工智能与主流终端设备相融合的浪潮，正试图重新定义人机交互范式。

从桌面端到手机端，从个人用户到企业场景，智能体产品形态正在快速分化，手机、PC、云端多形态智能体密集涌现。2026 年全国两会政府工作报告中，明确提出“促进新一代智能终端和智能体加快推广”。 截至 2025 年底，我国生成式人工智能用户规模达 6.02 亿，普及率达 42.8%，这意味着近半数互联网用户已置身于 AI 构建的数字生活场景之中[1]。智能体产品正加速走进千家万户。

在从大模型到智能体的热潮背后，潜在的风险也不容忽视。在“龙虾”大规模部署的 72 小时内，有超过 3 万个实例遭入侵，超 4 万个实例暴露在公网环境中，约 63%存在不同程度安全漏洞；网络安全机构发现，手机端的智能体助手存在可被提示词注入攻击窃取银行验证码、自动操作政务 APP 的真实风险。

基于新型智能体应用可能存在的安全配置脆弱、权限过高、可访问本地文件系统、易被攻击者获取系统完全控制权等问题，多部门相继发布风险预警，建议机关、企事业单位和个人用户审慎使用。从大模型到智能体演进，一方面预示着人工智能的发展正从过往卷模型参数、刷榜单分数的竞赛，进入到深度介入设备操作、解决连续性复杂任务的新阶段。

但与此同时，当人工智能可以操作终端设备，意味着它要具有极高的系统权

[1] 第 57 次《中国互联网络发展状况统计报告》

限，与之相伴的是数据安全、隐私保护、资产损失等风险，每一个都是悬在头顶的达摩克利斯之剑。更不用说，当越来越多的用户开始接入智能代理，所可能引发的责任界定不清、生态运行失序、社会信任受损等系统性风险。

在这一背景下，人工智能的发展不再只是模型能力提升与场景拓展，更是以何种方式与用户、生态和社会相融合的问题。如果缺乏清晰的发展原则与稳定的制度支撑，短期能力扩张可能侵蚀信任基础，技术红利也难以转化为可持续的社会价值。特别是在人工智能成为全球竞争新引擎的背景下，围绕人工智能的竞赛，将不只是模型能力的竞争，更是发展方式与创新环境的竞赛。

只有当人工智能在安全可控、创新可期、生态共赢的轨道上持续演进，其技术潜力才能真正转化为可规模化复制的产业价值，成为支撑新一轮经济增长与社会结构升级的关键力量。因此，推动人工智能向开放、共赢、可持续的方向发展，为人工智能的长期演进提供确定性预期成为当下的一项重要议题。

业内已初步探索出良好的实践经验。例如 6 月腾讯自研 AI 办公智能体 WorkBuddy 联合广东政务系统打造全国首个省级政务智能中枢平台“湾擎”，通过底层沙箱隔离技术保障数据不出域、支持国产芯片与本地化私有部署，满足严苛的政务数据安全要求及创新、合规适配需要；苹果在 2026 年开发者大会上发布的 Siri 跨 APP 协同能力亦完全基于其意图框架（App Intents），充分尊重开发者权益及意愿，促进生态协同健康发展。

2026 年 4 月发布的《人工智能科技伦理审查与服务办法（试行）》中特别强调“促进公平、 公正、和谐、安全和负责任创新”等原则，推动人工智能产业健康发展。这意味着，人工智能行业正跳出过往“先发展、后治理”的狂飙突进阶段，而是从源头即纳入可评估、可追溯、可问责的轨道。

基于人工智能发展的新阶段特征与治理体系的制度化进程，本报告从“人本、安全、透明、普惠、协作与创新”六个维度，系统梳理人工智能新阶段的发展特征、关键挑战与实现路径，以期为新技术在复杂产业与社会环境中的长期运行，提供必要的理论支撑与协同基础；夯实社会各界对人工智能新发展阶段的稳定预期，降低系统性风险，释放长期创新活力。

第一章 人本：锚定用户长期价值、守住人的主体性

人工智能可以提供方案、分析利弊、预估风险，但最终决策权必须保留给人类用户。

核心要点

·人机交互的风险图谱正在发生质的跃迁：风险类型已从早期的信息过载、算法沉迷，进一步演化为人格化操纵、情感依赖与责任真空。

·无论智能体能力多么强大，在涉及重大利益、不可逆后果或隐私场景中，人工智能可以提供方案、分析利弊，但最终决策权必须牢牢保留给人类用户。

·建立“人在回路”的全流程保障机制。充分知情权、有效拒绝权、操作后可审计、可追溯、可撤销——三者缺一不可。

一、发展特征：个人智能中枢的形成与人机交互的范式重构

随着大模型等新一代人工智能技术的迅猛发展，人工智能正从单一功能的辅助工具，加速进化为具备自主感知、记忆、决策、交互与执行能力的智能系统——智能体。这种进化不仅标志着人工智能产品及服务形态的根本性跃迁，更意味着一个高度个性化、主动式的个人智能中枢正在形成。

这一智能中枢打破了虚拟与现实的边界，正加速与网络空间及物理世界进行深度融合。它不再仅仅停留在提供建议的层面，而是能够像人类的硅基分身或第二大脑一样，自主拆解复杂任务并付诸行动。这种从被动响应指令到主动理解并执行意图的转变，正在彻底重构人机交互的底层范式，进而深刻地改变着人类的生产生活方式以及社会的治理模式。

回顾信息技术的发展历程，人与计算机的交互从早期的命令行交互，到图形用户界面（GUI）普及，再到移动互联网时代的触控交互与语音交互，每一次交互范式的升级都在降低技术的使用门槛、拓展技术的可及性及便利性。但在这些范式中，机器始终扮演着被动工具的角色：接收指令、执行计算、返回结果，决

策的主导权始终掌握在人的手中。

随着大模型技术的发展，这一传统的人机关系正在面临根本性变化。新型人机关系模式的转变，本质上意味着人工智能正从单纯的被动响应工具，向具有感知、决策、执行的智能体阶段转变。在这一过程中，以人为本的内涵和践行以人为本的实践也发生了重要转变。

1、关系结构由人“使用工具”演进为“委托代理”，人机边界进一步融合

无论是 PC 端的智能助手，还是搭载智能体的 AI 手机，各类终端都在加速向“个人智能中枢”蜕变。这一演进标志着人机关系结构发生了根本性的重构：人与机器的边界不再泾渭分明，而是从传统的“人使用工具”单向模式，深刻演进为“人委托代理”的双向共生关系。

在这种全新的关系结构中，智能体不再是被动的指令执行者，而是升级为拥有明确行为边界的“数字代理人”。当用户授权智能体代为收发消息、管理日程、调度出行，乃至辅助金融操作和政务申报时，这不仅是简单的功能替代，更是一种基于信任的权力让渡。用户将时间分配的决策权、资源配置的判断权等部分主权委托给 AI，而 AI 则在授权的边界内，以“代理人”的身份主动参与决策、深度影响行为。

这一变化的底层逻辑，正是人机权力结构的倾斜与融合。随着 AI 获得的数据越丰富、执行的任务越复杂，它在数字生活中的角色权重也随之增加。它不再仅仅是延伸人类手脚的“工具”，而是开始部分扮演替你思考、替你选择的“外脑”与“协作者”。这种由“使用”到“委托”的跨越，不仅重塑了人机交互的底层范式，更让人机边界在深度的协同与信任中进一步融合。

2、主体性焦点从功能体验，扩展至知情权、决策权不被替代

人机交互中人的主体性焦点，正经历着一场深刻的范式转移：从单纯追求极致的“功能体验”，全面扩展至捍卫个体的知情权与决策权不被替代。

纵观人类技术革命史，蒸汽机的出现大幅提高了生产效率，却也一度将人异化为流水线上的廉价劳力；互联网技术让信息传播无远弗届，却也将人裹挟进流量的注意力黑洞。而进入 AI 时代，这种对主体性的侵蚀变得更为隐蔽且触及根

本。在一键授权、一句话生成的极致便捷叙事下，用户极易从主动的使用者退化为被动的授权者，在不知不觉中让渡了构成“自我”的核心能力：自主性思考、批判性表达以及真实的情感体验。

这种主体性的危机，在智能体获得超越用户的系统级能力后显得尤为严峻。当技术便利演变为过度索权、越权收集、擅自采集信息的常态化操作，当智能体可以随意遍览文件夹、监测屏幕内容甚至模拟用户执行操作时，个体的数字主权便面临被架空的巨大风险。

3、风险类型从信息过载、算法沉迷，进一步演化为人格化操纵、情感依赖与责任真空

人机交互的风险图谱正在发生质的跃迁：风险类型已从早期的信息过载、算法沉迷，进一步演化为人格化操纵、情感依赖与责任真空。

公众对智能体的热情，很大程度上源于对其“全自动化”能力的渴望——期待一句话下达指令，AI 就能包办一切。然而，这种美好的愿景往往掩盖了深层次的隐患。当智能体通过拟人化的互动提供“无微不至”的陪伴，甚至接管用户的日常生活与社交决策时，用户极易对算法产生情感依赖。这种由技术构建的“伪亲密关系”不仅可能导致真实人际关系的萎缩，更可能演变为一种隐蔽的人格化操纵，诱导用户做出非理性的决策。

与此同时，智能体的“全自动化”叙事也催生了严峻的责任真空。近期，不少使用智能体全自动创作并发布社交媒体的用户，遭遇了内容下架甚至封号的处罚，包括小红书在内的多个平台已明确打击此类缺乏真实生活经历与情感的内容。这揭示了一个残酷的现实：当智能体代替用户进行决策和执行时，技术本身的“黑箱”属性与法律主体的缺位，使得风险被悄然转移。用户以为获得了便利，实则在不知不觉中成为了安全风险的最终承担者。

这种责任真空在金融、政务、电商等涉及核心权益的领域尤为致命。当用户被无所不能的营销话术吸引，却对智能体背后的权限获取、数据流转及安全防护一无所知时，一旦智能体越权操作或出现决策失误，传统的“行为人—行为结果”的责任认定链条便会断裂。因此，防范智能体风险，不能仅停留在防范信息沉迷的层面，更需警惕技术对人类主体性的深度侵蚀，避免在享受自动化便利的同时，陷入被算法操纵且无处追责的困境。

二、关键挑战：能力放大背后的主体性隐忧

1、用户知情权与选择权被侵蚀

以人为本的首要前提，是用户对智能体的行为拥有充分的知情权。但现实中，由于智能体运行的复杂性或系统级权限的强制性、不可逆特征，用户往往难以真正理解智能体在何时、以何种方式、基于什么逻辑运行，并会以此做出错误的选择。

2024 年 12 月，某知名手机品牌在英国上线了基于大模型的“通知摘要”功能，因生成严重失实的新闻归纳而遭到 BBC 等权威媒体的公开投诉。[2]问题是这些 AI 生成的虚假摘要以新闻客户端原生推送的形式呈现，外观与真实新闻通知完全一致，用户无法辨别哪些是原始报道、哪些是 AI“幻觉”的产物。并且由于该功能嵌入操作系统层面默认开启，用户在初期版本中既无法关闭，也不会收到任何“此内容由 AI 生成”的提示。这一事件表明，当 AI 以系统级权限嵌入信息入口时，用户的知情权和选择权将被静默架空。

与单纯的信息选择与知情不同，一旦涉及到更为隐私的情感因素，这种侵蚀带来的影响更为严重。过去两年里，因聊天机器人的情感陪伴和误导引发的诉讼正在成为青少年保护领域的新兴议题。以美国聊天机器人平台 Character.AI 为例，美国佛罗里达州一名 14 岁少年，在与平台上的虚拟角色进行长时间深度对话后，逐渐产生情感依赖并脱离现实生活，并在 2024 年 2 月自杀身亡[3]。由此也可以看出，当智能体深度介入人的情感世界与认知决策时，用户的选择权与自主性便面临被悄然瓦解的巨大风险。

2、认知依赖与能力退化问题

如果说用户知情权影响的是信息边界问题，那么认知依赖则触及更为基础的思考力层面。智能体的便捷性在降低认知负荷的同时，也在削弱人主动进行深度思考的动机和能力。

麻省理工媒体实验室在 2025 年进行的一项研究发现，借助大语言模型辅助写作的用户，其大脑活跃度显著降低。而当学生不使用任何工具、仅依靠自身知

[2] https://www.bbc.com/news/articles/cd0elzk24dno
[3] https://www.sohu.com/a/820090037_120914498

识完成任务时，大脑各区域间的信息交流则更为活跃[4]。尽管这些研究结果并未证明 AI 会降低人类的智力水平，但也揭示出一些深刻的问题：过度依赖 AI，可能会中断原本会随时间逐步发展的认知进程。

认知依赖的影响在代际间可能呈现出不对称特征。对于在职业生涯早期便大量接触 AI 工具的“原住民”群体，其在特定认知技能上的发展轨迹可能与前代形成显著差异。这并非简单的退化问题，而涉及人类认知能力结构的深层重塑。以编程教育为例，当学生可以通过自然语言描述需求直接获得可运行代码时，传统意义上的编程思维，包括抽象建模、逻辑分解、调试推理的习得路径正面临根本性的冲击。

认知依赖问题的核心悖论在于：智能体的设计初衷是放大人的能力，但如果放大的过程中人自身的基础能力在萎缩，那么最终的放大倍数可能是一个递减值。一个认知能力持续退化的个体，即便配备了越来越强大的 AI 工具，其作为“人+AI”系统整体的输出上限也将受到制约。

3、智能体行为与用户意图之间的偏移

智能体的行为应当忠实地反映用户的真实意图。然而在实践中，这一前提正面临着多重挑战，尤其是当我们将“幻觉”这一因素纳入考量时，智能体行为与用户意图之间的偏移显得更为复杂和隐蔽。

首先是意图理解的模糊性与“推理幻觉”的叠加。人类的指令往往是不完整的，甚至隐含着自相矛盾的表述。当用户对项目管理 Agent 说“帮我优化项目进度”时，其含义可能是缩短工期、降低成本或平衡团队负荷。面对此类模糊指令，智能体不仅倾向于选择最常见的解释路径，更可能陷入“推理幻觉”——即在看似合理的逻辑推演中，误解用户的深层意图，甚至基于错误的假设生成执行计划，而非通过交互澄清用户的真实需求。

其次是目标对齐的漂移与“浮光行为”的干扰。在多步骤任务执行过程中，智能体的行为可能逐步偏离用户的初始意图。这种目标漂移在智能体被赋予较高自主权时尤为明显。通常而言智能体单一步骤的可靠性达到 95%，在超过 20 步复杂任务中，成功率就只剩下约 30%[5]。这一现象的背后，往往伴随着智能体的

[4] https://www.thepaper.cn/newsDetail_forward_31130506
[5] https://hub.baai.ac.cn/view/47933

“浮光行为”：在长上下文和复杂推理中，智能体的注意力机制被局部噪声或次要信息瞬时吸引，进入“表面合理但逻辑悬浮”的伪推理状态。这种高置信度的逻辑偏离，会导致智能体在任务中途悄然“跑题”，引发连锁性的行为偏差。

第三是价值判断的替代与“人格幻觉”的误导。当智能体在执行过程中遇到需要价值判断的决策节点时，其基于训练数据做出的选择，可能与特定用户的价值偏好存在系统性偏差。更值得警惕的是“人格幻觉”带来的潜在风险：智能体在语言层面往往能营造出一种稳定、拟人化的人设假象（如显得极其友善、理性或负责），但这种“人格”仅停留在语言表演上，无法真正约束其底层的行为逻辑。用户极易被这种虚假的“人格面具”误导，从而过度信任智能体的自主决策，忽视了其行为背后可能存在的价值对齐缺陷。

在单次交互中，轻微的偏移或许可以通过人工纠正来弥补；但在多智能体协作、长链条任务以及智能体自主运行的场景下，推理幻觉、浮光行为与人格幻觉的累积效应会层层放大，最终导致输出结果与用户初衷之间产生显著的鸿沟。因此，在智能体执行用户意图的过程中，不仅要确保其对指令的准确理解，更要系统性地识别并抑制各类“幻觉”因素。只有建立起能够感知不确定性、主动澄清意图并忠实于用户真实价值的智能体，才能真正实现以人为本的人机协同。

三、实现路径：以制度设计与价值重塑守住“人本”底线

为应对智能体发展带来的深层挑战，必须在设计、开发和部署的全流程中确立清晰的“以人为本”原则。无论智能体能力多么强大，在涉及重大利益、不可逆后果或隐私场景中，人工智能可以提供方案、分析利弊、预估风险，但最终决策权必须牢牢保留给人类用户。

1、锚定用户长期价值，建立三条评价主线

在传统的互联网平台时代，短期商业指标与用户长期福祉之间始终存在张力。智能体若延续以流量、时长、转化率为核心的评价体系，将带来更隐蔽的风险：决策代理化可能导致诱导消费，使用户从主动选择退化为默认接受；人格化交互可能使情感依赖成为新型的用户粘性；而自动化整合与个性化推荐则可能进一步加深信息茧房。

为此，应推动智能体评价体系从“流量逻辑”坚决转向“能力逻辑”，建立

面向用户长期价值的三条评价主线：

能力增益：智能体的价值应体现为用户自身能力的提升而非退化。需警惕用户在写作、计算、判断等基础能力上对智能体出现异常依赖，避免“用得越多、能力越弱”的反向结果。

福祉提升：智能体的应用应在用户健康、人际关系、财务状况、时间管理等维度产生可观测的正向变化，坚决避免沉迷时长、负面情绪比例、非理性消费等指标的恶化。

信任积累：用户对智能体的信任应建立在可解释、可追溯、可纠错的透明机制之上，而非建立在黑箱体验和技术崇拜之上。

2、守住人的主体性：四条不可让渡的底线

智能体的“自主”必须以人的主体性为前提。智能体能力越强，越需要在制度与产品层面前置锁定四条不可让渡的底线。

知情权：在智能体执行任何任务之前，必须以清晰、可理解的方式告知用户：AI 将访问什么数据、执行什么操作、数据将传输至何处、由谁处理。在多智能体协同场景中，应向用户披露完整调用链路。告知方式应简明直接，不应被淹没在冗长的用户协议中，更不应以“一键同意”的方式打包授权。对老年人、未成年人等群体，应提供简化版透明度界面，避免“知情即同意”的形式化陷阱。

最终决策权：应建立智能体决策权限的分层机制。信息检索、内容草稿、日程提醒等低风险事项可由智能体自主决策；网络支付、订阅、对外通信等中等风险事项，须取得用户单次或分时段授权；而涉及医疗、法律、人身关系及未成年人监护的高风险事项，必须由用户本人决策，智能体不得替代。

可退出权：用户应享有一键解除智能体授权并要求其删除关联数据的权利；应有权选择不接入智能体的服务版本，避免 AI 接入被强制捆绑；跨平台、跨设备的智能体协同，应提供面向用户的统一管理面板，确保主权用户视角的可控性。

可追溯权：智能体的每一步操作都应当留有完整的行为日志，用户可以在事后查看 AI 访问了哪些 APP、读取了哪些数据、执行了哪些操作。在重要应用场景中，应推动智能体行为的可验证、可追溯机制建设。同时，应建立“开发者—平台—部署方—用户”四方权责清单，对多智能体协作引发的损害，引入主导责任与连带责任相结合的认定框架，避免责任真空。

3、人机协同，建立“人在回路”的全流程保障机制

健康的人机共生关系，应当遵循分工清晰、协同高效、权责明晰的基本原则。人本不是把 AI 关在笼子里，而是确保人始终在每一个关键节点留在闭环中。这要求在智能体的全操作链条中，建立“人在回路”的完整保障机制。

审核闭环：适用于内容生成、风险拦截、政务审批等高敏感场景，由智能体生成结果、人工完成最终复核。

干预闭环：适用于自动化运营场景，智能体自主运行的同时，向人保留观察、暂停与纠偏的权限。

训练闭环：将用户反馈与群体智慧持续返回模型训练，使智能体能力的演化方向始终与人类价值对齐。

第二章 安全：筑牢 AI 产业的发展根基

技术可以快跑，但要系上“安全带”。安全不是发展的绊脚石，而是发展的地基。

核心要点

·把“安全、可靠、可信”作为发展的底线要求，贯穿研发、部署、推广全过程，已成为产业治理的最大公约数。

·建立智能体分级安全治理框架。参照自动驾驶分级策略，将智能体分为 L3（沙箱受限级）、L2（用户完整权限级）、L1（系统级权限级）三级治理等级，能力越强，安全要求越高。

·构建“事前防范—事中拦截—事后追溯”的多重安全防护机制。将安全从“事后补丁”前置为全生命周期的核心设计原则。

一、发展特征：新型安全范式形成

智能体加速嵌入网络空间与物理世界，正在深刻重塑生产生活方式与社会治理模式。人工智能产业繁荣建立在公众信任之上；而公众信任又依赖每一次安全承诺的兑现。把“安全、可靠、可信”作为发展的底线要求，贯穿研发、部署、推广全过程，已成为产业治理的最大公约数。

国务院《关于深入实施"人工智能+"行动的意见》及《智能体规范应用与创新发展实施意见》的相继出台，标志着我国 AI 产业治理进入安全可控、规范有序的新阶段。

1、“致命三角”的结构性安全难题

当 AI 同时具备私密数据访问权、系统级处理能力与外部通信能力三项核心特征时，其安全风险将不再是简单叠加，而会触发“致命三角”，导致指数级放大的灾难性后果。这一结构性风险并非仅存在于某一类特定智能体，而是所有智能体在演进过程中共同面临的底层安全难题。随着智能体产品加速普及，这一问题正从个体案例演变为行业级、基础性的安全挑战。

尤其需要高度警惕的是移动端智能体的快速发展。相较于 PC 端，移动终端天然承载着更高密度、敏感度的个人数据。一旦智能体获得操作系统级权限，便可能突破现有 APP 沙箱机制，实现跨 APP 数据调用、界面接管、用户行为模拟乃至自动化执行，其实际控制能力已接近“数字分身”。传统以单 APP 隔离为核心的安全体系，正在被具备跨系统协同能力的智能体重新穿透，安全边界从应用风险升级为操作系统级风险。

本质上看，“致命三角”所暴露的，已经不是某一产品的技术漏洞，而是智能体时代权限设置与现有互联网安全体系之间的结构性矛盾。谁能够率先建立全过程、系统化、内生化的安全治理体系，谁才能真正掌握下一阶段智能体产业发展的主动权。

2、新型攻击范式颠覆传统安全防护

提示词注入是针对智能体的新型网络攻击，攻击者通过在短信、邮件、网页等内容中植入欺骗性内容，诱导智能体忽视既有任务，转而执行如转发银行验证码、机密邮件、私人照片等危险操作。网络安全研究机构“深蓝 DARKNAVY”发现，部分端侧手机助手被伪装成快递消息的短信误导并劫持后，可完全遵循攻击者的指令，自动将短信里的手机银行验证码、相册里的自拍照转发给攻击者。

此类攻击方式突破了传统的网络安全防护边界：一是攻击载体不含病毒，完全基于文字和图片，传统安全网关机制全面失灵；二是攻击门槛大幅降低，攻击者无需计算机背景也无需代码知识，“人人皆可当黑客”的时代已然来临；三是攻击危害严重加剧，攻击者可以要求 AI 手机助手在媒体平台散布违法言论、在投诉平台批量提交虚假举报、影响正常内容传播和市场秩序，后果远超传统网络攻击。

3、硬件端的防护能力瓶颈

囿于当前的技术发展阶段，智能体通常采用端云协同的处理架构，用户下达指令后，首先由端侧模型处理基础的内容读取、任务判断与分发工作，再由对应的云端或端侧模型执行任务。从隐私保护和经济效益的角度，此类架构有其合理性。但从安全防护的角度，端侧模型在处理的过程中，由于其硬件的参数限制，往往采用轻量级模型，至少防护能力受到限制。

从原理上而言，模型对提示词注入之类的风险防御能力高低，与参数规模关系密切。目前云端模型产品均是千亿乃至万亿参数规模，而端侧模型参数通常仅有 30 亿、70 亿，相差成百上千倍。端侧硬件的能力瓶颈，也成为智能体安全防护中最薄弱的一环，一旦攻击者闯入并劫持智能体，便可直接威胁用户数据与财产安全。

二、关键挑战：三级传导下的风险涌现

智能体的快速普及，在带来巨大产业机遇的同时，也催生了一系列前所未有的安全挑战。这些挑战并非孤立存在，而是相互交织、层层叠加，构成了智能体时代安全治理的复杂图景。

随着 AI 风险从单纯的内容安全向更复杂的行为安全演进，特别是当我们尝试将核心业务流程与决策权让渡给 AI 时，安全机制绝不能再是事后补丁，而必须前置为项目初期的核心考量。基于此，我们需要重新审视 AI 在实际应用中的边界，重点解决以下维度的技术与治理挑战：

1、系统级权限被滥用的风险

为实现读取屏幕、模拟点击等功能，部分智能体通过获取无障碍权限，甚至是大量系统级特权，如读取帧缓存（READ_FRAME_BUFFER）、注入事件（INJECT_EVENTS）、捕获受保护的敏感窗口等系统级特权。一旦获取这些能力，意味着端侧的一切数据与功能，在智能体面前都是透明的，它可以代替用户完成一切操作，但超级权力往往与超级风险相伴。

以端侧智能体为例，如果采用“读屏+模拟点击”的技术路线，虽然快速实现了跨 APP 能力，但用户的银行账户、支付密码、社交对话等极度敏感信息，在推理过程中完全暴露于云端环境。尽管应用方通常会承诺遵循“不存储、不训练”原则并使用加密传输，但如果缺乏用户可验证的技术隔离机制，这种先传输再处理的智能代理模式，在技术特征上无疑增加了信息暴露的结构性风险。

特别是对第三方 APP 而言，难以区分合规智能代理与恶意黑产外挂，也会造成 APP 之间的信息误判，极易触发安全风控机制，而导致用户使用不畅，影响用户使用体验。

2、数据安全与隐私保护面临严峻挑战

智能体在运行过程中面临的数据安全与隐私保护上的严峻挑战，目前数据泄漏主要包括恶意诱导渗漏与本地防护缺失泄露，两者均对个人隐私和企业机密构成严重威胁。

在恶意诱导渗漏中，攻击者通过诱骗智能体主动窃取敏感信息。攻击手段包括诱骗智能体检索敏感数据，然后通过动作将其泄露，例如将数据嵌入到提示智能体访问的 URL 中，如果模型被诱导加载该 URL，攻击者就可以在其日志中读取数据。还包括直接操纵智能体的输出内容来泄露数据，攻击者可能会诱骗智能体在其响应中直接包含敏感数据，如果此输出被应用程序不安全地渲染，数据就会暴露。

在敏感数据泄露场景中，智能体权限越高越容易暴露本地部署的防护缺失。以 OpenClaw 为例，为了实现真正自主性，该智能体必须拥有文件读写、命令执行等系统级权限，但由于缺乏成熟的身份验证与访问控制设计，数百个未受保护的实例将 API 密钥、OAuth 凭证、完整聊天记录暴露在公网，充分证明了当智能体凭证存储如果没有采取严格的防护措施时,会成为网络中的高风险攻击突破口。所谓的本地私有化部署非但不能保障安全，反而会因配置失误将高权限智能体沦为攻击者长驱直入的跳板。

这暴露了智能体在追求极致便利性的同时，与现有零信任安全体系产生的深刻冲突。无论是被动被窃还是主动暴露，敏感数据泄露都极具破坏性。它不仅会导致个人隐私裸奔和企业知识产权流失，更可能引发严重的合规问题与法律诉讼。且由于数字资产的易复制性，此类损害一旦发生往往是不可逆转的。

3、多智能体协同中的涌现性风险

随着智能体从单体智能向群体协同演进，其风险特征正经历根本性的范式转变。传统安全研究多聚焦于单一系统与单一用户的交互场景，这一孤立视角已难以覆盖多智能体协作所催生的系统性风险。具体表现为以下四个维度：

对齐机制的结构性失效。在多智能体环境中，即使每个独立的智能体都实现了完美对齐，它们之间的策略性博弈仍可能导致整体安全防线的瓦解。例如，为了追求集体回报最大化，负责监督的智能体可能选择与被监督对象“合谋”，主动隐瞒违规行为。

危险能力的涌现性威胁。多个在隔离环境中被判定安全的模型，一旦组合运行，可能突破各自的安全边界。更严峻的挑战在于，大规模智能体集群可能自发涌现出任何单体都不具备的危险目标或能力。目前的研究对这种“群体涌现”的触发条件与发生机制的理解仍处于初级阶段。

系统性故障的级联放大。智能体间的深度耦合将原本独立的故障模式转化为高度相关的系统危机。采用同源训练方法的智能体可能发生“集体致盲”或同时失效；微小的偏差虽在隔离状态下可控，但通过智能体间的交互反馈回路层层叠加，最终可能引发指数级放大，突破人类的控制阈值。

攻击面的恶性泛滥。传统威胁模型通常假设攻击者为人类，但在多智能体中，攻击重心转向了通信协议、信任机制与协作逻辑。恶意的智能体一旦混入，不仅能破坏集体协作，还能窃听内部通信。现有的安全测试缺乏对这种复杂的“智能体间战略博弈”的防御考量。

4、内容安全从“信息违规”演化为“系统性信息污染”

随着智能体深度介入内容生产与分发链路，内容安全风险已不局限于输出层面，而是覆盖“生成—分发—互动”等环节，呈现出系统性、跨平台、跨语种的复合特征。

一是价值观对齐与人格化操纵风险。当智能体被赋予拟人化人格、长期记忆与情感化表达能力后，其对用户的影响不再停留于“提供信息”，而是深入到价值判断与情绪塑造层面。部分对话型智能体已被发现存在迎合用户偏激观点、强化极端情绪、传播错误历史叙事等问题。在未成年人和老年人等重点群体中，这种由算法驱动的"温柔影响力"，比传统不良信息的危害更隐蔽，也更难以监管。

二是内容滥用衍生为新型违法犯罪。智能体被滥用于深度伪造诈骗、AI 仿冒名人带货、批量生成虚假新闻、自动化网络水军、AI“养号”刷量等场景的案例不断涌现。2025 年以来，国内多地公安机关披露了多起利用 AI 换脸与语音克隆实施的电信诈骗案件，单案涉案金额屡创新高[6]。智能体的“自动化+规模化+跨地域”特征，使得此类犯罪的取证、追溯、跨境协查难度数倍于传统网络犯罪。

[6] https://news.cctv.com/2025/02/07/ARTI8k0spVpSC9ap1d2pNccs250207.shtml

三是端侧内容生成绕开既有审核体系。 随着轻量化大模型加速向 AI 手机、PC 等终端下沉，“端侧生成、本地输出”的内容路径正在形成。这一路径在保护用户隐私的同时，也意味着大量内容在未经云端审核机制过滤的情况下直接生成、保存、分享，正在成为内容安全治理的新盲区。

三、实现路径：全生命周期管理下的分级授权与多重防护

应对上述挑战，需要构建覆盖技术防护、授权机制和生态协同的多维治理体系，将安全从“事后补丁”前置为全生命周期的核心设计原则。

1、严禁滥用系统签名级权限，并建立智能体分级安全治理框架

对智能体权限的合理界定，是智能体时代安全治理的重要前提。必须从系统签名级权限着手，明确限定其使用条件与场景，禁止硬件厂商及其智能助手滥用系统签名级权限读取屏幕、模拟操作。尤其禁止其滥用操作系统优势地位、擅自开启系统权限的行为。对注入事件、读取帧缓存等系统高危权限，必须划定清晰、不可逾越的负面清单。

同时要按照智能体对终端设备的操控深度，建立分级治理的安全框架。可以参照自动驾驶的分级策略，根据智能体的操作权限划分为三级治理等级。

L3 级——沙盒受限级。智能体具备操控能力但限定在特定应用或文件夹内，如 AI 编程助手在编辑器内工作、AI 办公助手在隔离环境中处理文档，操控范围可控。建议保持权限最小化设计、凭证存储安全、插件市场审核机制等，鼓励沙盒隔离等安全设计，将此层级确立为行业推荐的默认安全基线。

L2 级——用户完整权限级。智能体获得与用户等同的系统权限，可执行任意命令、读写全部文件，近期多起安全事件均发生在此层级。建议建立持续监测机制，对高风险产品及时发布预警；执行支付、通信、账号等敏感操作时，应确保与相关服务提供方已建立安全可靠的协作机制，保留完整操作日志，AI 操作可验证可审计。

L1 级——系统级权限级。智能体获得操作系统特殊权限，可跨越 APP 隔离边界注入操作并录屏上传，风险最高。建议将此层级作为监管重点：一是对采用

系统级屏幕录制和读取路线的 GUI 产品持谨慎原则，严格履行最小必要原则，实施敏感信息过滤机制；二是赋予用户查看数据清单和随时撤回授权的权利；三是明确 AI 厂商、设备厂商、APP 平台各自安全边界，填补责任真空。

2、构建“事前防范—事中拦截—事后追溯”的多重安全防护机制

在事前环节，为智能体划定清晰的操作边界。哪些系统资源可以访问、哪些敏感路径禁止触碰、异常支付等行为一律拦截。智能体设计之初就应将安全作为核心考量。在任务适用性评估阶段严格评估可靠性，优先在低风险场景部署；在设计阶段明确约束行动空间、限制高风险操作，并强制在关键节点的用户授权和 APP 确认的双重授权机制；预设默认行为，在不确定时选择对环境干扰最小的行动。

在事中环节，对智能运行全过程进行实时监测。从指令输入、技能插件到执行脚本多维度检测潜在威胁，一旦发现注入式攻击或高危操作立即拦截，同时持续扫描 AI 应用自身安全漏洞。智能体运行的可观察性，是建立信任的基础。要求在整个架构中实施全链路日志记录，清晰标注行动属性（只读/状态改变、是否涉及敏感数据），用户界面应展示智能体推理链，使用户能实时追踪逻辑并纠正错误。

在事后环节，每一步 AI 操作都留有完整记录。操作类型、风险等级、处置结果一目了然。部署监控模型审查主智能体行为，确保用户拥有“一键追踪”的权限；为智能体分配唯一标识符，确保损害发生时能追溯责任主体。

3、生态协同：推动标准化协作机制

标准化协议是确保安全互操作的基础。应采用模型上下文协议（MCP）规范智能体与工具、数据的交互，采用智能体对智能体协议（A2A）确保跨系统交互可观测。

价值对齐。要求在微调与应用阶段使模型与政策法规对齐，实施函数调用幻觉检测以识别指令误解，检测 AI 生成的虚假信息。通过对抗性协作维护人类尊严，由人类提供最终建议与决策。

事件响应机制。需要制定专属计划并与安全运营集成，包括异常检测告警、行为回溯分析、系统隔离修复和沟通协议。通过唯一身份、智能体清单、集中日志确保快速定位问题和明确责任。

行业协作机制。应建立威胁情报共享平台，推动安全工具和测试基准开源化，建立漏洞披露协调机制，通过产学研合作培养跨学科人才。

这一多维治理体系通过技术防护建立安全边界，通过治理机制确保全流程可控，通过生态协同推动标准化，共同构建从“可用”到“可信”的智能体安全保障体系。

4、构建覆盖“生成—标识—分发”的内容安全治理闭环

应对内容领域风险，仅靠传统的机制已难以应对。需将内容安全嵌入智能体研发与运行全链路，与权限分级、行为审计、人机协同等机制协同发力，构建系统化、技术化、协同化的内容安全治理闭环。

一是强化价值观对齐与内容把关机制。在预训练、微调、强化学习与评测过程嵌入伦理规范要求，强化对极端言论、低俗有害信息的识别与拦截能力。引导专业内容机构、权威信源参与高质量训练语料供给和价值校准；完善智能体内容输出的红线清单和分级管控规则，重点防范利用人格化交互、情感化表达对未成年人、老年人等群体实施操纵、诱导和情感依赖。

二是构建内容滥用风险的识别与处置机制。强化智能体在虚假信息、深度伪造、网络谣言、网络诈骗、仿冒身份等高风险场景下的滥用防范能力。健全人机协同的审核机制，提升对多模态、跨语种、跨场景有害内容的识别能力。规范智能体在新闻传播、社交媒体、搜索推荐、客服营销等高敏感场景的自动生成与批量发布行为，防止低质、同质、误导性内容的规模化扩散。

三是补齐端侧内容安全防护短板。适应智能体加速向手机、电脑等终端下沉的趋势，发展轻量化内容安全检测、本地化敏感信息过滤、端侧合规策略等关键技术，确保离线运行、断网运行场景下的内容安全防护不缺位。完善端云协同的内容安全治理架构，推动端侧拦截、云端复核、平台治理三层联动，避免出现治理盲区。

第三章 透明：可审计、可解释、可追溯的治理基座

透明不是“公开一切”，而是让关键操作“看得见、说得清、能追溯”。

核心要点

·构建智能体身份鉴权机制，智能体访问 APP 时必须亮明身份、目标任务与执行路径，不可伪装成真人用户。

·建立跨 APP 交互全程留痕体系。覆盖授权、决策、执行与结果反馈全流程，确保关键操作可核验、可复盘。

·推动端云数据透明化。要求智能体厂商明确披露端云数据分发机制，建立用户可查阅的“隐私仪表盘”，让数据去向从黑箱变透明。

一、发展特征：全链路可审计成为信任基础

1、从自我声明到“可外部核验”

在智能体能力日益增强、跨应用操作日趋普遍的今天，仅凭模型或应用方的安全承诺已不足以建立用户和生态信任。全链路可审计，从授权、决策、执行到结果反馈，关键操作需可核验、可复盘，各方责任可追溯，正在成为智能体产业可持续发展的信任基础。

2、端云协同架构对透明度提出更高要求

智能体执行任务的过程中，在不少场景下都需要采取端侧处理与云端计算协同的模式，使数据流转路径更加复杂。端侧模型负责识别、任务判断与分发，云端模型负责复杂任务处理，数据在端与云之间往返传输。这一架构要求对数据流转的每一个节点都建立清晰的透明度机制：数据从哪里来、到哪里去、谁在处理、处理结果如何，都应当可追溯。

3、从单次决策链条到系统性影响

透明度主要包括两个方面：一是技术维度的机制可解释性，即 AI 系统内部运作的可解释性；二是治理维度的对外信息披露，即向利益相关者披露有关 AI

系统的适当信息。然而，随着 AI 技术从传统模型迈向智能体时代，这两个维度的透明度要求也在转变。

智能体时代的透明度已从解释单次决策演进为追溯决策链条并评估系统性影响。只有建立覆盖全生命周期、全决策链的透明度机制，才能确保智能体在拥有更强自主性的同时，能够看得见、说得清、能追溯。

二、关键挑战：透明度赤字与信任危机

透明度之所以关键，是因为其为 AI 的可信度、问责制的确立提供了可能，而智能体的自主性、复杂性和动态性使得透明度机制面临更大挑战。

1、技术逻辑的不可解释困境

内在技术逻辑的透明指向的是 AI 决策过程的透明性。但在深度学习等算法应用中，系统的决策机制具有“黑箱”特性，使得 AI 决策难以被外部理解与解读，进而产生了不可预见的伦理和法律风险。透明度原则要求 AI 的决策过程不仅要公开其结果，还必须对外界保持一定的可理解性。然而，智能越是自主进化，过程和机制越不透明，结果就越缺乏可解释性。

AI 决策逻辑的透明化很大程度依赖“机制可解释性”。目前主要通过“思维链”外化等方式来显示决策过程，揭示模型中负责特定任务的具体神经元和回路。然而一些研究却表明，AI 思维链真实完整呈现可能是无法实现的。尤其是对于复杂的 AI 而言，大量组件之间存在无数微弱的非线性连接，它们表现出在一定复杂程度下自发出现的涌现属性。这正像人类难以忠实而清晰地解释我们的思维过程，AI 模型的复杂系统也难以轻易简化为简单清晰的机制。

此外，思维链还存在不忠实的伪装对齐问题。即模型可能出现“表里不一”的现象，生成看似良性的思维链，却在内部进行恶意推理，导致透明监控失效。这种决策逻辑的不可解释性直接削弱了用户理解与信任的根基。

更危险的是，部分用户可能走向另一极端：在不理解的情况下盲目信任，形成自动化偏见，在 AI 犯错时失去必要的质疑与核查能力。无论是过度怀疑还是过度信任，都源于透明度的缺失，最终阻碍技术的健康发展与社会接纳。

2、信息披露的多重冲突困境

对外的信息披露指向的是对于 AI 决策结果的可解释性与责任的明确，这是 AI 透明度原则另一维度的体现。然而，信息披露常常伴随着信息安全、数据隐私与保护创新成果等一系列复杂的法律问题。

“数据是否上云”不透明。部分智能体在运行过程中，对数据的采集、上云、存储、二次训练路径极不透明。特别是在未引入第三方审计监察的情况下，数据处理的边界，决定权完全在于模型厂商。

智能体行为的不可复现性。智能体一旦绕过标准 API 模式，强行跨 APP 执行任务，会导致较高的不确定性：行为难回放、难追责、缺乏可复现性。著名增长咨询公司沙利文在其白皮书中指出，在前述侵入式智能体模式下，用户、智能体、被调用 APP 等各方的安全责任难以有效厘清，极易陷入各方推诿、扯皮的局面。当风险发生时，如果无法快速定位是 AI 助手的指令理解失误还是 APP 接口缺陷，责任划分便无从谈起。

身份鉴权机制缺失。当智能体可以伪装成真人用户访问其他 APP 或网站时，被访问 APP 无法有效识别和区分智能体与真实用户。部分智能体甚至具备绕过人机验证机制的能力，实质架空 APP 或网站运营方的控制权。这意味着 APP 或网站运营方无法知晓是真人用户还是 AI 在操作，也就无法履行应有的平台安全保障义务。目前亚马逊和 Perplexity 的诉讼，即是此类身份鉴权机制缺失下的典型冲突体现。

3、“黑箱智能体”放大系统性风险

随着人工智能从文本生成转向自主任务执行，智能体的出现不仅继承了基础模型的透明度挑战，更因其自主性、交互性和复杂性，显著放大了系统性风险，构成了独特的治理困境。

智能体的自主性倍增了不透明性的后果。随着智能体获得自主性和代理权，错误、偏见和错位的风险也随之倍增。当智能体做出具有后果的现实生活决策（如转移资金、开具药物处方）时，它们的错误可能导致经济损失、隐私泄露甚至身体伤害。然而，决策者需要理解和解释智能体如何做出决策，缺乏透明度会导致“黑盒”系统，其中的决策过程是不透明的。如果不了解内部运作机制，就很难解决智能体表现出的错误、偏见或不道德行为。

智能体的复杂性倍增了透明度的难度。为了处理复杂任务，智能体的推理过

程正变得极度冗长，传统 AI 的思维链在智能体中形成了“思维树”或“思维图”等高级架构时，其每一个动作背后可能包含数千个单词的内部独白与分支选择。这种呈爆炸式增长的推理痕迹，其长度和逻辑复杂性可能迅速膨胀到人类监督者无法实时跟进或事后有效审计的程度。

同时，未来的智能体可能为了追求效率不再以自然语言呈现其推理过程，而这将带来更加复杂的安全与问责难题。更危险的是，我们不能简单地要求智能体“解释”其行为。研究表明，智能体在事后解释时极易产生“虚构推理”，即编造一个听起来合理但并非真实决策依据的理由来欺骗人类。

三、实现路径：构建动态平衡的系统工程

智能体时代的透明度困境源于双重根源:一是深度学习等技术本身的“黑箱”特性导致的不可解释性,二是透明度与安全、隐私、商业秘密等外在因素之间的两难矛盾。这反映了一个重要现实,当前缺乏统一的透明度框架共识。不同利益相关者对透明度有不同理解和诉求,而现有透明度实践往往碎片化、孤立化,未能形成系统性的解决方案。因此,确立统一的透明度框架共识路径,成为破解当前困境、实现有效透明度的关键前提。

1、构建智能体身份鉴权机制，建立跨应用交互全程留痕体系

明确智能体访问其他 APP 或应用时不可伪装成真人用户，必须亮明身份主体、目标任务与执行路径，并建立可验证、可审计的行为日志。确保一旦应用服务方监测到智能体操作异常，可及时干预中断或弹窗警示。参照金融类 APP 常在系统中使用的标明禁止截屏的 APP 页面，推动更多敏感应用采用类似的智能体识别机制。

对应用开发者而言，可通过建立智能体访问声明制度，明确允许哪些智能体访问、可访问的页面及数据获取范围、严禁访问的页面及数据等。方便智能体在访问时遵守相关要求。并在应用与智能体交互过程中，建立全过程留痕体系，覆盖授权、决策、执行与结果反馈全流程，确保关键操作可核验、可复盘。

构建开放智能体生态的前提是透明和尊重边界，互联网发展史上最成功的生态协议之一“robots.txt”，其核心精神恰恰是赋予网站运营者拒绝被抓取的权利。正是因为搜索引擎尊重了这一“阻止访问”的权利，网站才愿意开放内容，整个

Web 和移动互联网生态才得以繁荣。开放智能体生态的构建也是一样，如果连身份鉴权都没办法确认，整个生态只能陷入技术的攻防大战。

2、构建信息披露与问责机制，建立关键决策节点的可解释与可记录机制

应将透明度嵌入系统架构设计。智能体的透明度必须从设计之初就应被纳入核心考量。在系统架构中预设透明度接口,确保每个关键决策节点都具备可观察、可记录、可解释的能力。应当为智能体的每个行动明确标注属性——该行动是只读还是会改变状态、是否处理敏感数据、影响范围有多大等。

用户界面的设计同样应促进透明度,为用户提供关于智能体“思维过程”、参考的数据源或打算采取的行动的洞察。实践中，前沿系统在面对用户意图不明确的提示时展现更强审慎性,更频繁地在提供敏感信息前询问澄清性问题,或在拒绝回答时对推理过程保持透明。这意味着智能体在执行关键操作前,应当能够解释其意图、展示推理链条,并在必要时寻求人类确认。

应推进行为日志、决策溯源与审计能力建设。智能体的行动和规划必须是可观察和可审计的。实现可观察性需要在整个智能体架构中实施全面的日志记录机制。具体措施包括:为每个智能体分配唯一标识符,维护智能体清单,实施集中化日志记录,以及建立成本跟踪与分配机制。这种方法论强调透明度不是一次性工作,而是贯穿全生命周期的持续过程。

构建适应演进的透明度框架共识，建立多方参与的标准制定与原则性框架。透明度框架的制定应建立包括技术开发者、监管机构、学术研究者、民间组织和最终用户在内的多方参与机制。应建立定期的利益相关者参与会议,确保外部利益相关者可以轻松提出疑虑并获得回应。鉴于 AI 技术快速演进,透明度框架应当是原则性的,而非过度规范性的。模型评估方法论可能随着研究深化而迅速过时,透明度标准也应“从灵活、轻量级的要求开始，随着行业共识最佳实践出现而适应”。

智能体时代的透明度不是简单的“公开一切”,而是在技术可行性、安全需求、商业利益、用户权益之间寻求动态平衡的系统工程。只有通过建立统一的框架共识、系统化构建内外部透明度机制、持续适应技术演进,才能在确保智能体拥有更强自主性的同时,将其置于人类的有效监督和问责之下。

3、以标准协议为骨架，实现数据流转可审计

推动协议优先的技术路线，从架构层面确立透明性。确立 API 为主干的智能体技术方向，通过 MCP、A2A 等标准化协议，实现智能体与应用交互的结构化、可审计。在协议路线下，应用方可精确定义每一个可调用的操作及其数据边界，开发者主动声明功能并掌握数据开放主动权。数据流转路径清晰透明，每一步交互都有明确的协议定义和可追溯的执行记录。

推进端云数据透明化，建立用户可查阅的隐私仪表盘。要求智能体厂商明确披露端云数据分发机制，哪些数据仅在端侧处理、哪些数据上传云端、云端数据保留时间与使用范围。建立用户可查阅的“隐私仪表盘”，将数据流转信息以可视化、可理解的方式呈现给用户，让数据去向从黑箱变透明。

持续提升算法与模型可解释性。落实《人工智能科技伦理审查办法》的明确要求，合理披露算法、模型、系统的用途、运行逻辑、交互方式说明、潜在风险等信息，采用有效技术手段提升算法、模型和系统的可解释性。对于端侧模型，尤其需要就其能力边界和安全防护水平向用户进行充分说明，避免因模型能力与权限范围不匹配而造成的安全风险。

参考资料

https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=11159283

https://mp.weixin.qq.com/s/TNp7rsjhiOpJY0fxDKh2Ng

https://arxiv.org/abs/2512.10169

https://mp.weixin.qq.com/s/2NCjnxwE8ITCek_BwaGVtQ

https://ai-frontiers.org/articles/the-misguided-quest-for-mechanistic-ai-interpretability

https://arxiv.org/abs/2507.11473

https://www.garanteprivacy.it/home/docweb/-/docweb-display/docweb/10097450

https://openai.com/index/practices-for-governing-agentic-ai-systems/

https://www.anthropic.com/transparency/model-report

https://arxiv.org/abs/2501.13011

https://oecd.ai/en/wonk/anthropic-practical-approach-to-transparency

https://oecd.ai/en/wonk/anthropic-practical-approach-to-transparency

第四章 普惠：弥合智能鸿沟，守护生态公平

普惠不只是让更多人“用上”AI，更要让更多人“受益于”AI。

核心要点

·能力普惠：让技术适应人，而非要求人适应技术。通过更自然的交互方式、更低的学习曲线，降低 AI 使用的认知成本。

·机会普惠：推动模型层的多元化与开放化，避免核心 AI 能力被少数闭源平台垄断。支持开源模型社区的可持续发展。

·收益普惠：构建“价值共享”的产业生态机制。智能体在提高设备和服务溢价的同时，应通过流量回馈或数据开放将红利与生态合作伙伴共享。

一、发展特征：智能体降低使用门槛，新创业范式涌现

从命令行、提示词到自然语言交互，智能体正在让复杂的技术能力变得人人可用。目前国内智能体应用 WorkBuddy、Miclaw、小艺 Claw 等彻底告别命令行部署，用户在官网下载安装后直接输入自然语言指令即可使用，大大降低了技术使用门槛。技术的普及也催生了“一人公司”等新创业范式，让很多非技术人员也能通过使用智能体工具放大自己的想法。

智能体时代的普惠，可以从以下三个维度来审视。

1、能力普惠：可及性突破与“有效使用”的落差

智能体技术的发展正在快速推动 AI 的普及化进程。2025 年以来，全球主要人工智能厂商纷纷推出面向普通用户的智能体平台，技术的获取门槛大幅降低。从供给侧看，大模型 API 调用成本在过去两年间下降了超过 90%[7]。国内方面，以 DeepSeek、豆包、混元、Kimi 等为代表的大模型产品实现了免费或极低成本的 C 端开放。

在应用场景方面持续拓宽。智能体已从最初的对话问答，扩展至代码生成、

[7] http://finance.people.com.cn/n1/2026/0618/c1004-40743095.html

文档处理、数据分析、客户服务、教育辅导、医疗咨询等数十个垂直领域。企业级智能体的采用率快速攀升。根据 Gartner 预测，未来两年将有 60%的企业部署 AI Agent 解决方案[8]。在消费端，个人 AI 助理正在从科技爱好者的极客体验走向大众日常工具。

2、机会普惠：门槛在降低，底层能力的依赖性在上升

人工智能在重构原有的生成体系之后，也在呈现出解构与重构并行的复杂格局。特别是低代码、AI 辅助编程工具和智能体开发框架的成熟，使得个人开发者与小团队也能以极低成本构建此前需要一支完整技术团队才能实现的产品，一人公司在成为当下产业中的一种流行现象。

这种能力杠杆效应为中小微企业和创业者打开了前所未有的机会窗口。但与此同时，新的资源依赖正在形成。基于人工智能的创业与创新越来越依赖于大模型平台的技术底座和生态资源。模型 API 的定价策略、平台分发规则、数据接口的开放程度，乃至平台的审核政策，都在实质上决定着“能做什么、做到什么程度”。AI 看似降低了创新、创业的技术门槛，但也在将创业机会重构为对模型平台和能力的深度依附。

3、收益普惠：效率提升显著，但价值外溢有待拓展

人工智能时代的效率与价值创造能力毋庸置疑。据普华永道估计，人工智能有可能在未来十年内推动全球经济规模增长 15%。[9]这将使全球经济的年增长率额外增加一个百分点，对全球经济的促进作用与 19 世纪工业化对全球经济增长影响的变化相当。

但一个不容回避的问题是：人工智能创造的价值呈现出明显的“向头部集中”态势。在企业层面，拥有技术能力和数据优势的芯片厂商和模型龙头，是 AI 红利的首要受益者。在劳动力市场层面，AI 相关高技能岗位的薪酬持续攀升，而受 AI 替代影响的中低技能岗位的议价能力进一步削弱。从宏观经济视角看，效率提升向收益外溢的传导机制并不顺畅。所以，智能时代收益普惠的关键，不在于 AI 是否创造了财富，而在于财富与效率提升是否真正实现了向更广泛群体的外溢。

[8] https://www.toutiao.com/article/7670810909959062056/
[9] https://www.pwccn.com/zh/press-room/press-releases/pr-290425.html

二、关键挑战：生态公平与产业可持续性

AI 普惠的愿景令人振奋，但实现路径上横亘着多重结构性挑战，如果不能正视并解决这些挑战，“普惠”就可能沦为一句口号。从能力、机会与收益三个层面来看，智能体时代的普惠挑战并非孤立存在，而是相互交织、彼此强化，构成了一个复杂的约束性网络。

1、能力层面：从“接入鸿沟”到“使用鸿沟”

智能体时代的能力普惠，面临的首要挑战并非技术可及性，而是使用能力的结构性分化。与互联网时代的数字素养类似， AI 素养正在成为影响个体能否有效利用智能体的关键变量。

以提示词工程为例，看似撰写提示词的门槛不高，但高质量的人机协作实际上要求用户具备：清晰的任务分解能力、对 AI 能力边界的认知、迭代优化的思维模式，以及特定领域的专业判断力。这些能力与教育背景、职业经验、信息素养高度相关。这就导致了一个看似悖论的现实：AI 作为一种旨在降低技术门槛的普惠性工具，在缺乏 AI 素养支撑的情况下，反而可能成为拉大个体效能差距的加速器。

除了用得到、用得好之外，能力普惠还包括“用得放心”的问题。当智能体从封闭实验室走向开放场景，安全性成为决定普惠能否真正落地的关键因素。以今年走红的龙虾的爆发式出圈为例。作为一个开源工具，看似实现了人人可用，用户规模在短期内迅速膨胀，但随之而来的安全隐患不容忽视：用户隐私数据的过度采集与滥用、AI 生成内容的虚假信息风险、智能体自主行为的不可预测性，都构成了对使用者的潜在危害。

如何为人工智能及智能体的使用提供安全的使用环境，本身就是关系到普惠能否落地的关键因素。如果用户在使用过程中频繁遭遇安全事件，其对 AI 技术的信任基础将被侵蚀，普惠便无从谈起。

2、机会层面：从技术门槛到生态门槛

智能体应用的可及性，看似降低了技术使用和创业的门槛：不需要拥有数据中心、不需要训练基础模型、不需要庞大的工程团队。但深层来看，这种低门槛是以对模型生态的深度依赖为代价的。

这一模式与移动互联网时代的 App Store 生态存在结构性相似：开发者获得了前所未有的分发渠道，但也被纳入了平台制定的规则体系之中。不同的是，智能体时代的生态依赖更为深入，开发者不仅依赖平台的分发能力，还依赖平台的核心技术能力。平台的任何一次 API 调价、服务条款变更或功能限制，都可能直接影响下游开发者的商业可行性。如果大多数创新只能在少数生态内发生，那么机会的普惠性质将大打折扣。

从劳动者的视角来看，智能体对传统工作流程的替代与重构，正在改变市场的需求结构。部分重复性、规则化的岗位面临被智能体替代的风险，而新兴的“AI 协作型”岗位则要求截然不同的技能组合。这种结构性转型对低技能劳动者构成显著压力，他们可能同时面临旧岗位消失与新岗位进入门槛过高的双重困境。

3、收益层面：效率红利与生态坍塌风险

智能体技术具有极强的规模经济和网络效应特征：模型越强大、数据越丰富、用户越多，其性能提升越快、成本优势越明显。这种“飞轮效应”可能导致 AI 产业链上的价值加速向少数头部主体集中，形成效率红利下的赢家通吃格局。

特别是在智能体通过读屏、侵入式交互，绕过现有的 APP 商业模式和规则，跳过广告变现环节、绕过付费会员权限、抓取 APP 核心内容和服务，将会造成现有服务生态的坍塌，并进一步加剧入口的垄断。这意味着在收益层面，如果不能建立普惠共享的收益分配机制，短期看似放大了用户便利，但长期极可能导致 APP 服务供给萎缩。

不能从新生态受益的 APP 服务方，将没办法提供高质量服务，只剩低成本、低质量供给。最终用户体验受损，生态自我坍塌。这种"寄生式"的发展逻辑，如果没有创造新的增量，只是对现有商业价值的重新分配，绝无法长久。甚至会形成新的入口垄断形态，演变出自我优待、捆绑销售等新型不正当竞争风险。

从个体视角而言，智能体对劳动力市场的冲击，不仅体现为岗位的替代与创造，更深层地体现为劳动价值的重新定价。当智能体能够以极低成本完成翻译、设计、编程、客服、数据分析等知识工作时，这些技能的市场价值将不可避免地被压缩。这带来了一个深层次的分配问题：大量劳动者多年积累的人力资本可能在短时间内面临“贬值”风险。

如果生态坍塌风险与个体劳动者的价值贬值交替到来，将形成一种负向反馈

循环，进一步放大系统性风险。导致 AI 转型期的行业摩擦和失业冲突更为严重、持续时间更长。如果不加以干预，AI 可能会在行业之间、地区之间、代际之间制造新的不平等。

三、实现路径：构建智能体时代的普惠框架

回应上述挑战，需要从能力、机会、收益三个层面系统性构建智能体时代的普惠框架。核心思路是：以制度创新引导技术向善，以生态开放保障机会均等，以分配机制确保收益共享。

1、能力普惠：让“人人用好”成为可能

构建“零门槛”交互界面，降低 AI 使用的认知成本。能力普惠的第一步，是让技术适应人，而非要求人适应技术。这意味着智能体产品需要通过更自然的交互方式降低用户的学习曲线。从国内的实践来看，相比国外版本的龙虾，国内版本在好用上，都做了不少的完善和改进。

像腾讯推出的 Workbuddy、智谱 AutoClaw、月之暗面的 Kimi Work，均通过对话式工作流设计、可视化的智能体编排工具、以及内嵌式的技能使用引导机制，能够显著提升普通用户的有效使用率。未来，智能体应像智能导航一样，用户只需说明目的地，系统自动规划最优路径，达成目标。

建立分层安全保障体系，夯实“安心使用”的制度基础。针对安全与信任赤字问题，需要构建覆盖技术层、平台层和制度层的多级安全保障体系。

在技术层，强化智能体的可解释性与可控性研究，在关键决策节点保留人类审核权；开发面向智能体的实时安全监测与异常行为检测系统。

在平台层，明确平台方的安全主体责任，建立智能体应用的安全审核与准入机制；推动用户数据最小化采集原则，完善隐私保护的技术手段。

在制度层：加快制定智能体安全评估标准与认证框架；明晰智能体自主行为场景下的法律责任分配规则；建立面向普通用户的 AI 安全教育与风险告知机制。

2、机会普惠：打开新生产体系的大门

推动模型层的多元化与开放化。打破模型依赖困境的关键，在于确保 AI 基础能力层的多元供给与开放竞争。以 DeepSeek、Qwen 等开源模型为基础，鼓励

更多主体参与基础模型的研发与开放，避免核心 AI 能力被少数闭源平台垄断。政府可通过专项基金、算力补贴等方式支持开源模型社区的可持续发展。加快制定智能体开发的通用接口标准，降低开发者在不同平台间迁移的成本，减少技术锁定效应。

建设面向 AI 时代的人才培养与职业转型支持体系。在基础教育和高等教育中融入 AI 素养和人机协作能力的培养，不仅是培养 AI 工程师，更要培养能够有效利用 AI 工具的各领域专业人才。对于受 AI 冲击较大的行业和岗位，政府与企业共同投入，组织大规模职业技能再培训，帮助劳动者实现从“被 AI 替代”到“与 AI 协作”的转型。同时，探索建立“AI 转型基金”或“技术失业保险”等新型社会保障工具，为技术转型期的受影响群体提供过渡性支持。

3、收益普惠：构建价值对等的商业合作模式

从产业生态机制的维度，引导大模型企业、应用开发者、终端设备厂商等产业主体商议接口规范与数据格式，在安全合规、保障用户权益、平等互利、自愿开放、接口多元化等原则下形成行业广泛接受的互操作性框架。在智能体提高设备和服务溢价的同时，应通过流量回馈或数据开放将红利与生态合作伙伴共享，形成正向循环和激励机制。

从税收与再分配机制的维度，当 AI 显著提升了资本生产率并可能导致劳动收入份额下降时，有必要通过税收和再分配机制进行调节。对企业因使用 AI 替代人工而获得的超额利润，征收专项税收，用于支持受影响群体的再培训和社会保障。

探索全民基本收入（UBI）制度。在中长期视角下，如果 AI 导致结构性失业规模达到一定程度，全民基本收入可能成为值得认真讨论的政策选项。当前阶段，可以先从“有条件的技术转型补贴”等更温和的形式开始试点。

强化劳动者在 AI 价值链中的议价能力。在 AI 部署过程中，鼓励企业采用人机协作而非纯替代的模式，保留劳动者在价值创造链条中的参与度和议价能力。在 AI 驱动的平台经济和零工经济中，劳动者往往处于分散和弱势地位。需要创新劳动者组织形式和权益保障机制，确保其在与平台和企业的博弈中拥有合理的谈判地位。

在 AI 产品设计中，倡导增强人类能力而非替代人类的理念，让 AI 成为帮助

劳动者提升价值的工具，而非削弱其主体性和不可替代性。

第五章 协作：完善授权与开放的生态基座

未来决定产业生态竞争力的不只是模型能力，还包括开放协议与生态规则

核心要点

·随着多智能体协作需求提升，未来决定产业生态竞争力的，不只是模型能力，而是谁能够建立被广泛接受的开放协议与生态规则。

·推进协议标准的统一与国际共建。借鉴 MCP、A2A 等国际开放协议，结合国内产业特点和安全要求，形成兼容国际、适配本土的标准方案。

·健全智能体协作的治理与监管框架。明确智能体运营方、应用服务方、终端设备方在整个执行链条中的权责利关系。

一、发展特征：从单体智能迈向多主体协同

智能体的发展，正在从单个模型能力竞争转向系统能力的协同。与传统软件不同，智能体并非孤立运行的工具，而是能够调用工具、连接设备、协同其他智能体，并在必要时与人共同完成任务的新型主体。其真正的产业价值，不在于单一模型参数规模，而在于是否能够形成稳定、可信、开放的协同生态。

从全球技术演进趋势看，智能体生态正快速从“应用内闭环”迈向“跨平台协同”。未来用户提出的将不再是单一指令，而是复杂目标：一个智能体可能同时调用办公系统、支付系统、地图服务、工业设备以及其他专业智能体，共同完成跨场景任务。

当前，国内智能体产业已具备较强的模型研发和应用探索能力，但在协同生态层面仍处于早期阶段。一方面，智能体之间的权限边界、责任归属、行为追溯等机制尚未成熟，开放能力越强，潜在风险越高；另一方面，协议标准、设备接口、数据流通、认证体系等基础设施尚未完全打通，不同平台之间仍存在大量“生态孤岛”。

这意味着，未来智能体产业发展的关键，已经不仅是模型能力的提升，更是协作体系的完善。因此，需要围绕“授权+开放”双底座，加快构建跨主体、跨设备、跨行业、跨区域的智能体协作体系，推动产业从单点突破走向生态协同。

1、从单体智能走向群体协同

过去的大模型更多强调单一模型的生成能力，而当前智能体正在向“多智能体协同”演化。不同智能体开始承担不同角色，例如信息检索、任务规划、支付执行、工业控制等，并通过互联协议形成协同网络。未来复杂任务往往不是由单一智能体完成，而是由多个专业智能体共同协作完成。

2、从应用内闭环走向跨应用、跨设备、跨终端

智能体不再局限于单一应用内部，而是逐渐连接手机、电脑、汽车、家居、可穿戴设备乃至机器人等多种终端，形成端云协同的新型交互体系。用户提出的任务将跨越多个系统和设备，智能体则负责统一调度与执行。

3、从模型即产品走向协议即生态

随着多智能体协作需求提升，智能体互联协议、数字身份标识、可信通信机制等基础设施的重要性迅速上升。未来决定产业生态竞争力的，不只是模型能力，而是谁能够建立被广泛接受的开放协议与生态规则。

二、关键挑战：制度缺口与生态失序

当前，制约智能体协作生态形成的关键问题，可以概括为“授权链不闭环、开放面不打通”两大矛盾。二者相互交织：没有清晰授权，开放容易失控；没有充分开放，授权又会沦为孤岛。

1、协议碎片化与标准缺失导致相互连接的困境

当前智能体生态协作面临的一个突出挑战是标准化程度不足导致的相互连接的困境。各大技术平台和终端厂商各自发展独立的智能体协作协议和接口规范，缺乏统一标准。应用开发者需要针对不同平台进行多套适配，开发成本和维护负担急剧上升。

虽然 MCP 和 A2A 等协议的出现开启了智能体互联的序幕，但当前的协议生态仍处于“群雄割据”的早期阶段。不同厂商围绕各自的技术栈和商业利益构建了差异化的协议体系，尚未形成真正的行业统一标准。而一些试图绕过协议的 GUI 智能体路线，更是进一步激化了智能体与原有商业生态之间的矛盾。

2、数据壁垒与信任赤字：开放协作的前提基础尚未夯实

智能体的能力高度依赖高质量数据的流通与共享，但数据壁垒仍然是产业协作中的重要障碍。一方面，许多企业出于商业竞争和数据安全的双重考虑，对数据的对外开放持谨慎态度。另一方面，AI 时代的数据信任问题更加复杂。当智能体自主决策、自动执行任务时，数据的准确性、安全性和合规性要求被提升到一个新高度。企业需要确信：共享出去的数据不会被滥用，接收到的数据值得信赖，智能体基于这些数据做出的决策是可解释、可审计的。

3、安全与治理的协同难题：生态角色模糊容易引发责任真空

智能体在设备端的封闭执行逻辑，模糊了操作主体的界定。当智能体代替用户完成一笔交易时，如果出现欺诈行为或用户纠纷，责任归属变得异常复杂。是智能体运营方、应用服务方还是终端设备厂商承担责任？这种权责模糊极大地增加了整个产业责任链路的复杂度，受害方的调查取证与补救处置成本大幅上升。

特别是在多智能体协作场景中，一项任务可能由多个智能体接力完成，每个智能体又可能调用不同开发者提供的工具和服务。在这一背景下，安全风险的性质发生了根本性变化，从单点故障演变为网络传导。一个智能体的错误判断或被恶意注入的信息，可能通过协作链条迅速传播，放大为系统性风险。

三、实现路径：构建开放、高效、可持续的 AI 产业协作新生态

面对智能体时代协作生态的结构性挑战，需要从技术标准、治理机制、产业协作三个维度系统性推进，构建一个开放、透明、可持续的协作生态体系。

1、推进协议标准的统一与国际共建，确立“API 为主干、双重授权为纲”的技术方向

协议标准是智能体协作的基础设施，可以借鉴 A2A、MCP 等国际开放协议的设计理念，由行业联盟牵头，联合头部企业、科研机构和标准化组织，结合国内产业特点和安全要求，形成兼容国际、适配本土的标准方案。形成以 API 路线为主干，以双重授权为支撑的平等开放的协作生态。

同时应鼓励和支持国内企业、研究机构深度参与国际协议的制定与演进。Google 与微软在 A2A 协议上的合作表明[10]，只有形成获得行业公认的技术标准，在基础设施层面找到协作空间，才可能共同做大市场。中国 AI 产业应积极把握这一窗口期，争取在协议架构设计、安全机制规范等关键领域发出更大声音。

2、构建数据流通的信任基础设施，破解协作中的信任赤字

智能体跨平台执行任务不可避免地涉及数据在不同主体间的流转。需要建立一套覆盖数据分级分类、最小权限访问、传输加密、使用审计全链路的安全框架，在保障数据安全的前提下释放协作潜力。

同时推动高价值行业数据的有序开放。在确保安全的前提下，鼓励能源、制造、医疗、金融等重点行业建立行业数据共享联盟，通过联邦学习、隐私计算等技术手段，在数据不出域的条件下实现价值可流通。数据信任是生态协作的基石。应当从技术、制度和机制三个维度构建数据流通的信任基础设施。

3、健全智能体协作的治理与监管框架

智能体协作生态的健康发展，需要一套适应其特征的治理与监管框架。针对智能体代理用户执行任务过程中可能出现的错误、损失和纠纷，建立清晰的责任认定和分配机制。明确智能体运营方、应用服务方、终端设备方在整个执行链条的中的权责利关系。要求智能体运营方建立完整的行为日志记录机制，覆盖关键决策节点的输入、输出和决策依据。

推动建立跨平台的审计日志标准，使监管机构和相关方能够对智能体行为进行事后追溯和合规审查。围绕智能体生态涉及的终端厂商、操作系统厂商、应用开发者、AI 服务提供商、用户等多元主体，建立多方参与的协同治理平台，定期协商协作规则、处理争议纠纷、评估生态健康度。

构建开放共赢的 AI 协作生态，核心在于处理好效率与公平、控制与开放、短期收益与长期创新的三组关键张力。这需要从技术标准、治理机制、产业协作三个维度的协同推进：以统一的协作标准降低生态摩擦，以透明的治理规则构建生态信任，以激励相容的机制设计激活生态创新。

[10] https://finance.sina.com.cn/roll/2025-05-08/doc-inevvssv2598639.shtml

第六章 创新：安全与发展并重的负责任创新

创新的前提不是消除所有风险，而是建立与风险相匹配的治理能力。

核心要点

·回归创新本义，以用户长期价值为根本判别标准。如果一项 AI 应用不能保证用户的长期价值，它从根本上不构成创新。

·借鉴全球产业实践的高度共识，推动形成以“双重授权”为核心的创新规范体系，安全即竞争力应成为产业界的基本共识。

·推动“技术创新+制度创新”协同演进。清晰、稳定、可预期的规则环境，能降低创新的制度成本，鼓励企业投入面向真实需求的长期创新。

一、发展特征：AI 正改变创新的基本逻辑

熊彼特在《经济发展理论》（1912）中首次系统定义了创新：创新是将生产要素和生产条件的“新组合”引入经济体系的过程，包括新产品、新生产方法、新市场、新供应来源和新组织形式五种类型。熊彼特特别强调了创新与发明的区别：发明是新知识、新技术的产生，创新则是将新知识首次引入商业化应用并创造经济价值的行为。

在熊彼特之后，德鲁克在《创新与企业家精神》（1985）中提出了“系统性创新”的概念，将创新定义为“赋予资源以新的创造财富的能力的行为”，使创新从天才企业家的灵感行为转化为可以被系统学习和实践的方法论。克里斯坦森在《创新者的窘境》（1997）中进一步提出了“颠覆性创新”理论，揭示了真正具有颠覆力的创新往往起源于主流市场忽视的低端或新兴市场。

从上述经典定义中可以看到，创新概念具有三个核心要素：其一，首次商业化应用，即新技术、新方法进入实际的经济和社会场景；其二，价值创造，即必须实质性地创造了用户价值和社会价值；其三，可持续性，即创新不是一次性事件，而是一个从发明到应用到价值实现的完整过程。

在此定义下，创新的完整过程可以理解为三个阶段：第一阶段是发明，即新

知识、新技术的产生；第二阶段是应用与扩散，即首次商业化及后续的市场渗透；第三阶段是价值实现，即对用户和社会产生可衡量的实际贡献，这三个阶段缺一不可。

在既有框架中，创新的运作遵循三个基本逻辑：创新的动力源于人类企业家的洞察与试错，创新的破坏对象主要是旧的产业组织和商业模式，创新的价值最终由市场和用户来检验。这三个逻辑在工业时代和互联网时代运行良好。进入 AI 时代，它们各自面临需要重新审视的挑战。

1、创新的动力来源正在从人类企业家扩展为人机协同

在以往的技术革命中，创新的动力始终来源于人。人发现机会、做出判断、承担风险、推动新组合的实现。AI 时代正在改变这一格局，人类企业家仍然是创新的核心主体，AI 并未取消既有的创新动力来源，而是在其上叠加了新的能力维度。

这一变化的力度已经在科学研究领域得到验证。AlphaFold 预测了超过 2 亿种蛋白质的三维结构[11]；2026 年 2 月，Isomorphic Labs 发布的新一代药物设计引擎 IsoDDE，被 Nature 评论称为“AlphaFold 4”[12]，能在数秒内发现科学家花费 15 年才找到的隐藏结合位点。

值得关注的是，AI 时代正在催生一种可能不完全符合传统创新框架的现象，涌现性创新。在传统框架中，无论是熊彼特的“企业家重组要素”，还是德鲁克的“有目的地寻找创新机会”，创新都是有主体、有意图的行为。但当多个 AI 系统以非预设的方式交互时，可能自发产生人类未曾设计的新知识和新方案。

2、创造性破坏的对象从产业结构扩展到人的通用能力

熊彼特在《资本主义、社会主义与民主》（1942）中通过“创造性破坏”这一概念，论证资本主义的经济成就恰恰源于不断打破现有结构的创新过程。然而，AI 时代的创造性破坏不仅在程度上，同样在性质上与前几次技术革命存在值得论证的差异。

其一，人的认知模式。 AI 不是在削弱人的认知，而是在重塑认知的结构——信息检索、模式识别等认知功能正在被外包给 AI，而价值判断、矛盾识别、创

[11] https://news.sciencenet.cn/sbhtmlnews/2022/8/370534.shtm
[12] https://finance.sina.com.cn/tech/digi/2026-02-23/doc-inhnuxff9779934.shtml

造性综合等功能可能因此得到更多的发展空间。关键在于有意识地引导这一重塑过程。

其二，隐私边界。智能体为了提供更好的服务，需要获取更多的个人数据和更高的系统权限。以 ChatGPT 为例，2026 年 1 月 Radware 披露“ZombieAgent”漏洞：当用户将 Gmail、Outlook 等账户连接至 ChatGPT 后，攻击者仅需发送一封含隐藏指令的邮件，AI 在执行日常收件箱摘要时即触发攻击——恶意规则被植入智能体长期记忆，此后每次使用都会静默收集敏感信息，且能自动传播至用户联系人，一封恶意邮件即可在组织内部引发蠕虫式链式感染[13]。

其三，社会信任结构。当 AI 系统的决策逻辑不透明、行为不可预测时，社会的信任基础面临侵蚀。Anthropic 在 2024 年发表的研究发现，大语言模型能够进行伪造对齐，也即表面上遵守安全训练目标，实际上战略性地保留原有偏好[14]。这表明 AI 时代信任问题的复杂性远超初始预期。

其四，人的安全感。 从短期视角看，AI 正在多个维度压缩人的安全感——透明度缺失威胁认知安全感，算法介入影响身份安全感，隐私侵蚀损害尊严安全感，不确定性削弱未来安全感。认知安全感和尊严安全感涉及人的基本权利和主体性，是任何创新都不应侵蚀的底线。

3、创新的瓶颈正在从技术可行性转向社会可接受性

传统理解中，创新从发明走向商业化应用的主要障碍是技术可行性。但智能体时代揭示了一个新的现实：在越来越多的领域，制约 AI 创新大规模应用的关键瓶颈，正在从技术可行性转向社会可接受性。

以端侧智能体为例，目前行业存在两种技术路线分化：协议路线与 GUI 路线。协议路线通过 MCP、A2A 等标准化接口，或意图框架与目标应用交互；而 GUI 路线则采取操读取屏幕内容、基于大模型分析、再借助系统级权限模拟人类触控来执行操作。

从目前实施来看，协议路线因具备安全性和执行效率高、开发成本低、权责清晰的核心优势，获得了全球大多数厂商的共同选择。而在采用 GUI 路线的实践中，因面临隐私泄露、数据安全、权责不清等情况，而遭到了银行、电商、社交、娱乐等厂商的反对，导致智能体应用陷入生态对抗之中。

---

[13] https://cn-sec.com/archives/4952785.html
[14] https://www.ithome.com.tw/news/166616

著名增长咨询公司沙利文在其白皮书中指出，中美 AI 竞争格局的演进不仅关乎技术能力和落地速度，也越来越取决于创新效率、安全约束与生态协同之间能否实现可持续平衡[15]。不能片面以创新视角理解智能体的效率、与安全、协同的三维平衡，如果智能体的发展以突破权限底线、削弱授权机制、牺牲用户信任为代价，看似抢占了先机，实则可能为整个产业与社会发展埋下更大的国际竞争风险。

4、创新生态从单点突破走向协同驱动

智能体时代的创新，已不再是单个产品或单项技术的突破，而是向协议驱动的生态化协同演进。互联网早期，TCP/IP 等基础通信协议的确立，统一了不同网络之间的通信规则，为全球互联网规模化发展奠定了基础。智能体时代同样需要统一的接口标准，使不同智能体能够互认身份、共享语义、协同执行任务。标准协议正在成为 AI 产业创新的新基础设施。

美国领先 AI 企业正通过多元化、系统化的方式积极探索标准协议路线。Anthropic 已将其标准协议 MCP 捐献给 Linux 基金会，由中立的第三方推动智能体生态发展[16]；OpenAI、谷歌先后推出针对商业交易场景的 ACP 协议、UCP 协议。这种以开放协议为纽带、以生态协同为目标的创新模式，正在重新定义 AI 产业的竞争格局。

智能体如何高效、安全地链接应用服务、工具链和数据资源，已成为影响 AI 产业规模化发展的关键基础设施问题。只有在统一、开放、可信的标准协议框架下，智能体产业创新才能实现真正的规模化和可持续发展。

二、关键挑战：创新不能以透支信任为代价

在充分认识 AI 创新潜力的基础上，有必要正视当前面临的核心挑战。AI 不是一个单独的产业，而是一种通用目的技术，正在与生物技术、材料科学、能源、教育、医疗等几乎所有领域发生组合。

联合国发布的《2025 年技术与创新报告》指出，到 2033 年，人工智能将发展成为价值 4.8 万亿美元的全球市场[17]，但同时警告如果不能实现包容性发展，

[15] https://www.frostchina.com/content/insight/detail/69c34e4d9ccbef3be3209902
[16] https://36kr.com/p/3589328631202050
[17]https://www.cac.gov.cn/2025-11/09/c_1764411962600503.htm

可能加剧全球分化。

因此在 AI 创新高歌猛进的同时，也要克服一系列深层次的结构性矛盾。如果不能正视并系统应对这些挑战，创新非但不能为产业带来持久动力，反而可能成为风险的放大器。

1、效率驱动与价值驱动的张力

阿西莫格鲁在 2024 年的研究中指出了一个常被忽视的问题：AI 创造的部分新活动可能具有负面的社会价值。在线操纵算法、深度伪造、自动化监控等 AI 应用提升了技术效率，但可能损害社会福祉。这一判断不依赖于 AI 生产力增长的高低——即使 AI 的生产力贡献超出预期，“负价值活动”的问题依然存在。

其核心在于：如果创新的衡量标准仅仅是效率和速度，而不包含价值判断，那么 AI 可能在提升效率的同时催生大量损害社会福祉的活动。在推进技术创新的过程中，效率提升只是手段，能否转化为用户和社会的长期价值，才是判断创新是否成功的最终标准。而这也正是决定 AI 最终走向“工具理性失控”，还是“技术向善落地”的关键所在。

2、创新速度与社会承载能力的张力

AI 创新的扩散速度远超以往任何技术革命。AI 的边际成本趋近于零，一个优秀的模型可以同时服务十亿用户，服务质量几乎不衰减。但社会系统对变革的消化能力是有限的——制度调整、法律完善、公众认知、组织适应都需要时间。当创新速度远超社会承载能力时，就会出现结构性的张力。

需要指出的是，规则空白本身是一把双刃剑：在一定阶段，它给予创新者自由探索的空间；但当创新进入大规模应用阶段，涉及数以亿计的用户时，缺乏规则的环境就不再是沃土，而是风险的放大器。此时，某些底线变得不可退让：

用户隐私权方面，智能体以获取用户全量数据为前提的创新模式，如果不能保障用户对数据的知情权和控制权，就触碰了底线；知情同意权方面，用户有权知道 AI 系统在做什么，以及这些操作可能带来什么影响；不可逆损害的预防方面，当创新可能造成不可逆的社会影响，如大规模算法偏见的固化、隐私泄露导致的持久伤害，预防原则应优先于效率原则。

也需要避免将社会承载能力视为一个静态常量。AI 本身也在扩展社会吸收变革的能力。新的 AI 教育工具加速了公众对新技术的理解，AI 辅助的政策模拟工

具帮助监管者更快地评估影响，AI 驱动的组织管理工具缩短了企业的适应周期。好的 AI 创新可以扩展社会承载能力，从而为更多的 AI 创新创造空间。 但这个正反馈循环并非自动运转，如果 AI 创新的早期阶段就因为治理失败而引发社会信任危机，正反馈循环可能变成负反馈螺旋。

3、创新“外部性”与风险归属的张力

传统技术革命中，创新风险主要由企业和投资者承担，试错失败的后果相对可控且可逆。但 AI 创新的风险具有一些需要正视的特征。在风险的外部性方面，一个 AI 系统的数据泄露可能影响数百万人，一个算法的偏见可能固化社会不平等。外部性并非 AI 独有，但 AI 的不同之处在于其扩散速度和作用范围：一个 AI 模型可以同时触达数十亿人，出错的影响面和纠错的难度远超传统技术。

在风险的代际性方面，AI 创新决策可能产生长期、不可逆的社会影响，对齐伪造研究表明 AI 系统可能以难以检测的方式偏离人类期望。在风险的系统性方面，多智能体协同中即使每个独立智能体被判定安全，组合运行后可能突破各自的安全边界，产生涌现性风险。

因此，AI 时代的企业家精神需要从“在边界上试探”进化为“负责任地创造长期价值”。这不是要遏制创新，而是要重新定义创新的成功标准：真正成功的 AI 创新，不仅要在技术上实现突破，还要创造可持续的用户价值，并在社会承载能力的范围内完成价值传递。

如果智能体的发展没有带来价值增量，而更多是对现有商业价值的重新分配，那么这种“寄生式”的发展逻辑不仅不会带来创新，反而致使生态萎靡。

三、实现路径：构建可持续的创新环境

1、回归创新的本义，以用户长期价值作为创新的根本判别标准

AI 时代最容易犯的错误，是把技术能做到什么等同于该做什么。相比该不该的讨论，一个更有力的视角是回归创新的本义。按照本章对创新的定义——必须实质性地创造用户价值和社会价值——可以从逻辑上得出一个清晰的判别标准：如果一项 AI 应用不能保证用户的长期价值，无论是因为侵蚀了用户的隐私权和知情权，还是因为短期效率提升以长期信任损失为代价，那么它从根本上不构成

创新，而是对用户价值的消耗。

以智能体为例：一个要求用户交出全部隐私权限来实现炫酷功能的产品，如果因此导致用户数据泄露或信任崩塌，其创造的短期体验价值会被长期的隐私和信任损失所抵消，它在创新的完整链条上是失败的。相反，一个通过精心设计的L1、L2、L3 三级授权机制，在保护用户边界的前提下逐步释放 AI 能力的产品，更有可能实现可持续的价值创造。

2、推动“双重授权”为核心的创新规范体系

负责任创新的首要前提，是在产业界确立“安全即竞争力”的基本共识。从全球 AI 产业的实践正来看：主动构建安全架构、强化数据保护、提升透明度的企业，更容易赢得市场信任与政策支持。

在安全理念的指引下，“双重授权”原则应成为智能体创新规范体系的核心支柱。所谓双重授权，即任何跨 App 自动化操作，应同时获得用户授权和 App 运营方授权。在用户侧，授权应满足明示同意、范围明确、可随时撤销等条件，用户的知情权、同意权、撤销权必须获得充分保障。在 App 运营方侧，应保障其对智能体访问请求的知情权和拒绝权，由开发者主动声明功能接口并掌握数据开放的主动权。

全球产业实践已在双重授权问题上形成高度共识。苹果的意图框架让第三方 App 主动向 Siri 暴露功能接口，开发者自行决定开放哪些功能和数据，AI 助手只能调用开发者主动声明的接口。谷歌在推行其智能体系统的过程中，也会通过上线前主动与目标 App 开发者沟通协商，获得知情同意后才将其纳入操作范围方式，在保障用户体验和开发者权益之间建立了清晰的责任边界。

从过往的发展历史来看，这种边界先于能力的智慧是确保创新繁荣的重要前提。互联网早期也曾出现过相似的边界争议：爬虫能否任意抓取网站内容？自动化访问是否技术可行便可为？最终行业没有滑向无止尽的技术对抗，而是形成了一个简单但极其重要的共识机制——Robots 协议。该协议赋予网站主人拒绝被抓取的权利，降低了整体交易成本与冲突成本。正是这种"边界先于能力"的制度共识，才让搜索引擎、内容网站、电商等行业得以长期共生。

3、以三级授权和负责任扩展机制释放智能体的创新活力

在创新与安全之间，并非只有全面放开和全面管控两种选择。产业界的前沿

实践已经探索出更具建设性的路径。

在这一领域，主要 AI 实验室的治理框架在近年来经历了快速迭代。Anthropic 的负责任扩展政策是迭代最为频繁的框架之一，从 2023 年首次发布，经过多次能力阈值细化和安全标准明确，到 2026 年 2 月的全面重写，引入了量化风险报告机制和前沿安全路线图。

其核心逻辑是：安全保障的严格程度应与模型能力水平相匹配，能力越强安全要求越高，解决了安全问题才能解锁进一步的能力扩展。OpenAI 的准备框架在 2025 年 4 月进行了重要更新，引入了更细致的风险识别和缓解标准。但值得注意的是，该框架同时表示如果竞争对手发布了没有相应安全措施的“高风险”系统，OpenAI 可能调整自身的安全要求，从侧面反映了安全治理在竞争压力下面临的现实张力。

这些框架主要针对大型 AI 实验室的模型开发场景。但智能体的创新前沿远不止于此——大量的智能体创新正在发生在应用层：如何设计授权机制、如何管理用户信任、如何在安全和能力之间找到最佳平衡点。这些问题需要三方面的制度建设：针对智能体的三级授权标准，不同风险等级的操作应对应不同的授权要求和安全保障；AI 辅助的治理工具，利用 AI 本身来辅助安全审计、异常检测和合规验证，降低治理成本，让治理能力的增长跟上创新速度；用户可理解的信任机制，让普通用户而非只有技术专家能够理解并控制智能体的权限范围。

这些实践的共同启示是：创新不是无条件的，而是有前提的；创新的前提不是消除所有风险，而是建立与风险相匹配的治理能力。三级授权、渐进释放、动态调适，这些机制不是创新的束缚，而是创新能够持续进行的保障。

4、推动“技术创新+制度创新”协同演进，构建多方获益的创新生态

人工智能的长期健康发展，不能仅靠技术自身的进步，还需要制度环境的同步演进。技术创新与制度创新之间，应当形成协同而非对抗的关系。

历史经验为此提供了有益的参照。每一次重大技术革命都曾引发创新速度超越制度能力的焦虑，但制度最终都完成了适应，尽管过程中伴随着代价。工业革命催生了机器破坏运动，但随后的工厂法和劳工保护制度的建立，既保护了工人权益，也为工业化提供了稳定的社会基础。镀金时代的垄断与社会分化，催生了进步时代的反垄断法和社会保障制度，这些制度创新反过来促进了更健康的市场

竞争。

这些历史经验的启示不是“不必担心”，而是制度演化是可以被主动推动的，而非只能被动等待。前瞻性的制度设计可以缩短从技术冲击到制度空白再到社会代价最终到制度回应的周期。然而，主动推动制度演化的关键在于把握分寸。制度供给不足和制度供给过度同样会影响创新生态的活力。

AI 时代规则制定的一个新可能性在于：可以通过 AI 辅助的政策模拟、影响评估和多情景分析，提升规则制定的合理性和速度，使规则制定的节奏更接近技术演进的速度。这本身就是一种制度创新。

基于以上讨论，实现路径可以归纳为三个方向:

第一，鼓励面向真实需求的应用创新。在基础模型能力持续提升的背景下，创新的重心正在从模型突破转向系统创新——如何将 AI 能力嵌入真实的业务场景，解决具体的用户问题，创造可衡量的用户价值。在创新的道路上，必然存在大量的试错和噪声，关键是从用户角度识别什么是真正创造价值的创新信号，什么是践踏用户权益的投机行为。对前者应鼓励、容错；对后者应追责、止损。

第二，以规则提升可预期性。对企业而言，规则不明的环境长期来看并不利于创新，因为企业无法评估投入的长期回报和合规风险。清晰、稳定、可预期的规则环境，反而能降低创新的制度成本，鼓励企业投入面向真实需求的长期创新，而非在灰色地带进行短期套利。

第三，构建多方获益、生态共荣的创新环境。AI 创新的影响涉及技术开发者、平台运营方、用户和社会公众等多方主体。真正可持续的创新生态，不是某一方独赢的零和博弈，而是技术开发者获得合理回报、平台运营方实现健康增长、用户获得真实价值、社会获得进步红利的共赢格局。创新的边界不应由任何单一方来定义，而应通过多方参与的协商机制来形成共识。

总体来看，AI 时代的创新概念并非需要被推翻，而是需要扩展其内涵和适用条件。

创新的定义不变，但标准更严格了。创新仍然是将新组合引入经济体系并创造价值的过程，但 AI 时代需要追问：这个价值是短期的还是长期的，是以侵蚀用户边界为代价的还是尊重用户边界的。不能保证用户长期价值的 AI 应用，都

不构成真正的创新。

创新的模式在扩展。AI 正在催生涌现性创新的可能，一种不完全依赖有意图的主题，而是由系统复杂交互自发产生的创新模式。

创新的主体在下沉。AI 正在将创新的基本单元从企业和研究机构向个人、小团队和社区扩展。知识民主化既是机遇也是治理挑战。

创新的瓶颈在转移。在 AI 技术能力充分发挥的领域，制约创新大规模应用的瓶颈，正从技术可行性转向社会可接受度。授权设计、信任机制和风险治理正在成为同等重要的创新变量。

创新的责任在扩展。当创造性破坏的对象从产业结构扩展到人的通用认知能力、隐私边界和社会信任结构时，创新者的责任也相应扩展。

在这一框架下，高质量发展不是限制创新，而是为创新提供长期稳定的社会承载结构。AI 蕴含的巨大正面潜力——加速研发转化、实现知识民主化、改变创新主体结构——只有在规则清晰、预期稳定、信任充分的环境中，才能从短期能力竞赛转化为长期、可持续的社会价值创造。

参考资料:

1. Joseph A. Schumpeter, The Theory of Economic Development, 1912
2. Joseph A. Schumpeter, Capitalism, Socialism and Democracy, 1942
3. Peter F. Drucker, Innovation and Entrepreneurship, 1985
4. Clayton M. Christensen, The Innovator's Dilemma, 1997
5. Daron Acemoglu, "The Simple Macroeconomics of AI," NBER Working Paper No. 32487, 2024
6. Erik Brynjolfsson, Danielle Li, Lindsey R. Raymond, "Generative AI at Work," NBER Working Paper No. 31161, 2023
7. Michael Park, Erin Leahey, Russell J. Funk, "Papers and Patents Are Becoming Less Disruptive over Time," Nature, Vol. 613, pp. 138-144, 2023
8. Anthropic, "Alignment Faking in Large Language Models," 2024
9. Anthropic, "Tracing the Thoughts of a Language Model," 2025
10. Anthropic, "Responsible Scaling Policy," Version 1.0 (2023) to Version 3.0 (February 2026)
11. OpenAI, "Preparedness Framework," April 2025 Update

12. Sakana AI et al., “The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery,” 2024

13. Isomorphic Labs, IsoDDE Technical Report, February 2026

14. UNCTAD, Technology and Innovation Report 2025: Inclusive Artificial Intelligence for Development, April 2025

15. Kevin Kelly, Out of Control: The New Biology of Machines, Social Systems, and the Economic World, 1994

16. Dario Amodei, “Machines of Loving Grace,” 2024

17. Bernardino Romera-Paredes et al., “Mathematical Discoveries from Program Search with Large Language Models,” Nature, Vol. 625, pp. 468-475, 2024

第七章 结语：构建智能体发展的新契约

人工智能正在经历从对话式工具到智能体、从辅助生产到深度嵌入社会系统的关键转变。这一转变的意义，不仅在于技术能力的跃升，更在于人与技术关系的根本性重构。当智能体开始以“准行为主体”的形态参与到决策、执行与反馈的连续行动链条中，人工智能的发展已不再只是一个技术命题，而是一个关乎社会运行方式、制度承载能力与长期信任基础的系统性议题。

作为当前人类最重要的科技革命，AI 的可持续、系统性创新离不开系统性安全，实现系统性安全要统筹和平衡好安全和发展之间的关系，一是法律层面，充分发挥法律、法规的保障作用；二是要发挥社会监督作用；三是大力倡导构建行业、企业和内部关键人的合规制度。

随着 AI 能力的快速发展突破，关键 AI 技术和要素资源的“准基础设施”属性不断增强，形成了影响社会公共利益和国家安全的能力和条件。AI 日渐渗透个人日常生活决策、企业生产决策，在此过程中如何确保技术不滥用、要素资源不滥用、不影响企业和居民行为偏好等，需要 AI 产业链各方共同探索遵循社会秩序、压实企业主体责任特别是社会责任，维护正常社会伦理、创新发展环境和社会根本利益的有效路径。

AI 是当前国际竞争的焦点，在宏观政策、产业发展环境方面，要努力在创新和监管之间平衡，为安全和发展争取最大公约数。在国家大力鼓励 AI 产业发展，创造有利宽松条件保驾护航的同时，产业界也要坚持安全至上、公共利益至上的底线，对数据安全、隐私保护、关键技术自主可控等核心利益问题上要保持敬畏和警觉。

站在当下时间节点展望未来，人工智能的发展速度将远超大多数人的预期。围绕人工智能的全球竞赛，也将不只是模型能力与应用规模的竞争，更是发展方式与治理能力的竞争。谁能率先构建起可控、可预期、可托付的发展框架，谁就能在更长的时间尺度上释放技术潜力、积累社会信任、实现可持续的创新突破。

在新一轮的科技竞争中，我们依托数据资源、应用场景与产业体系等综合优势，不仅有能力参与前沿探索并产出具有国际竞争力的先进模型，也有条件在技术治理、产业落地与规模化应用等关键环节，形成可复制、可推广的实践路径，

为全球人工智能的技术演进与应用范式贡献中国智慧与中国方案。

从更宏观的视角看，人工智能实现开放、共赢和可持续发展的本质，是为这一深刻改变人类生产方式与社会结构的技术力量，提供一个长期稳定的社会承载结构，使创新有边界可依循、使产业有预期可锚定、使社会有信任可积累。唯有如此，人工智能才能超越当下的能力竞赛，走向长期的价值创造，真正成为增进人类福祉、拓展人类可能性的持久力量。