

智能体时代授权的逻辑限度与法律建构

李汶龙

(浙江大学 光华法学院,浙江杭州 310088)

摘要: 以大模型为基础的智能体正在重塑人机协作与任务执行方式,其在开放环境中自主拆解目标、调用工具并跨系统执行操作,使传统基于单一行为、单次授权的逻辑面临挑战。我国《智能体规范应用与创新实施意见》已将用户授权范围、决策边界、行为追溯以及用户最终决策权置于智能体治理的核心,但现有规范有待从原则层面进行细化,对复杂授权关系中的行为归属与责任分配进行系统回应。本文提出应超越以自然语言为中心的授权表达,转向以机器可执行的权限边界与能力约束为核心的结构化治理机制。在分析智能体场景下授权载体不充分、授权表达不清晰、授权主体不显明以及授权后果不可控等困境基础上,提出以授权辅助、权限控制、外观管理、公平交互与责任承担为核心的智能体授权治理框架。责任配置亦需从用户授权转向服务提供者的结构性责任,通过可追溯性、可中断性、偏离控制与授权链条管理等制度工具,实现从授权证明到责任映射的转型。只有在授权表达机制与责任体系同时重构的前提下,智能体经济方可成为可控可持续的制度基础设施。

关键词: 授权; 代理; 权限控制; 责任分配; 智能体经济

中图分类号: D922.16

文献标志码: A

文章编号: 1009-8003(2026)04-0066-14

一、问题的提出

随着生成式人工智能向智能体(agent)形态演进,人工智能系统正从信息生成工具转变为能够执行任务、调用资源并直接作用于现实环境的行动主体。^①新一代智能体不仅能够生成文本或图像,还可以读写文件、调用API、跨应用操作,甚至完成支付、签约与资源调度等具有法律与经济后果的行为。智能体不仅在任务执行方式上突破了传统“输入—输出”的工具逻辑,在决策链条、行为空间与环境交互模式上呈现出前所未有的自主性与复杂性。^②与生成式人工智能相比,智能体更强调在开放环境中自主拆解任务、调用多种工具、跨平台执行操作并对外产生法律关系变更,这一技术路径使其同时成为生产流程的执行主体、用户接口以及潜在的交易与合约入口。^③这一转变中,人工智能不再只是辅助决策的“工具”,逐渐成为参与乃至主导行动过程的执行节点,从而将原本停留在信息层面的风险,推向财产、合同与责任分配等法律领域。与此同时,人机关系亦发生深刻变化。传统模型下,由人作出决策、机器负责执行的结构正在被打破,取而代之的是一种持续互动的人机协同模式:用户以模糊指令表达需求,智能体通过多轮对话进行拆解、试探与重构,不断修正、生成行动路径。这一过程中,决策不再是一次性完成,

收稿日期: 2026-05-05

作者简介: 李汶龙(1989-),男,法学博士,浙江大学光华法学院“百人计划”研究员、数字法治研究院研究员、博士生导师,主要研究方向: 数字法学、前沿人工智能监管、人工智能立法。

①参见 Stryker, Cole. "What is agentic AI?" <https://www.ibm.com/think/topics/agentic-ai>.

②参见 Krishnan, Naveen. "Ai agents: Evolution, architecture, and real-world applications." arXiv preprint arXiv: 2503.12687 (2025). <https://arxiv.org/abs/2503.12687>.

③参见 Aksoy, Pinar Çağlayan. "AI as Agents". In DiMatteo, Larry A., Cristina Poncibò, and Michel Cannarsa, eds. The Cambridge handbook of artificial intelligence: global perspectives on law and ethics. Cambridge University Press, 2022, pp.146-160.

而是在交互中逐步生成,行为结果往往难以被还原为某一明确的授权或单一意思表示。^①

智能体经济正在成为平台、终端厂商与大模型企业共同关注的下一代产业入口。自豆包手机助手短暂展示其跨应用操作、读屏执行与任务编排能力后,很快因底层权限、安全边界和平台生态难题而下市。^②但这一事件并没有阻断产业发展方向,反而暴露出智能体经济进入消费终端之前必须跨过的制度门槛。据报道,豆包手机助手正式版项目已在推进,新机型计划在2026年推出。与此同时,国外厂商也在加速布局。OpenAI 被报道正在推进面向智能体时代的硬件或手机合作方案,^③Meta、Google 等企业亦在开发能够主动完成日常任务、深度嵌入既有生态的智能体助手。由此可见,大模型、工具调用、跨应用编排、自然语言交互和终端侧权限整合,已经使智能体具备从信息生成工具转向任务执行者、交易入口乃至资源调度者的条件。智能体时代已经临近,只是还不确定以何种制度形态表现出来。

在此背景下,中国相继出台多项人工智能相关政策文件。2025年8月,国务院发布《关于深入实施“人工智能+”行动的意见》,从产业与社会应用层面作出系统部署,强调要推动人工智能在重点行业与关键场景中深度赋能。^④2026年5月,由国家网信办、国家发展改革委、工业和信息化部联合印发的《智能体规范应用与创新发展实施意见》(下称“《实施意见》”)把智能体作为独立对象加以规制与引导,成为当前我国智能体治理最为核心的政策背景与制度锚点。^⑤《实施意见》直接回应了智能体可决策、可行动、可持续运行的技术特征,将用户授权范围、决策边界、行为追溯机制以及用户最终决策权等核心问题置于规则设计的核心,在授权结构与行为控制方面提出了多个具有前瞻性的要求,至少可以概括为决策权限厘定、行为管控与制度衔接三重关切。上述政策导向可以看到,中国在智能体治理上已经初步形成围绕“授权—决策—行为”的结构化问题意识。然而,现有规范更多停留在原则性要求与技术治理倡导层面,对于智能体在复杂授权关系中的法律地位、具体责任边界以及传统民商法、数字法、平台治理框架下既有授权制度的延展适用,仍有大量问题有待进一步展开系统分析。

智能体经济能否从低风险、边缘化应用走向金融、医疗、政务、企业管理等高价值场景,关键并不只是模型能力是否足够强大,还在于社会是否敢于将真实资源、真实流程和真实法律效果交由智能体处理。这里存在是两层相互关联但性质不同的授权问题。第一层是用户对智能体进入其系统环境、读取数据和内容并代为行动的授权,即用户是否允许智能体访问屏幕、文件、邮箱、通讯录、支付账户、日程系统以及其他个人数字环境,并在此基础上完成任务。第二层是智能体调用第三方应用、接口或平台服务的授权,即用户的一次同意能否支持智能体进入特定APP、调用平台功能、绕过或重组原有交互界面,并对第三方平台、服务提供者或其他用户产生影响。前者更多涉及用户与智能体之间的权限开放、信任关系和行为归责,后者则进一步牵涉平台边界、数据接口、生态竞争与跨主体责任分配。

当前背景下,授权的重要性被进一步凸显。没有清晰、可验证、可审查的授权结构,智能体行为的合法性基础、权限边界与责任归属便难以被有效界定。与此同时,授权本身的局限也更加明显。即便形式上存在用户同意或授权,法律也难以继续简单假定行为结果当然等同于授权意图,更不能将授权理解为免除责任的一般性理由。基于此,本文结合多智能体系统的运行机理与典型风险场景,系统考察传统授权逻辑在智能体语境中的适用困境及其重构路径,重点关注三个层面的问题:其一,有效授权在智能体场景中应当具备何种构成要素;其二,授权边界如何在自然语言指令、系统权限配置和动态任务执行之间被形成、呈现与调整;其三,当智能体行为超出用户预期或授权范围时,相关后果应如何在用户、开发者、部署者、平台和第三方服务提供者之间进行责任分配。

①参见 Hacker, Philipp and Kai Ebert. Attributing Behavior, Knowledge, and Intent—Principles, Rules, and Questions for Software and AI Law. in: Christiane Wendehorst (Ed.), AI in Contracting, 2026 forthcoming, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5423894.

②参见陈天昊《端侧智能体的监管定位、风险结构与规范构建》,载《东方法学》2026年第1期。

③参见 Bonifield, Stevie. OpenAI is reportedly launching a phone for ChatGPT. 5 May 2026, <https://www.theverge.com/ai-artificial-intelligence/924063/openai-phone-rumors-2027-ming-chi-kuo>.

④参见国务院《关于深入实施“人工智能+”行动的意见》(国发〔2025〕11号)。

⑤参见国家网信办、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》。

二、智能体及其风险

理解智能体为何会对既有授权逻辑提出挑战,需对其技术机理和特征作简要勾勒。本文将智能体定义为以人工智能模型为核心、以目标或指令为牵引、能够在有限人类监督下感知环境、规划任务、调用工具并连续执行操作的准自主系统。与生成式人工智能体相比,智能体的本质特征在于从内容生成走向行动执行,能进入由数据、账户、设备、平台、接口和第三方服务构成的行动环境,并通过自动化或半自动化操作产生财产、合同、隐私、竞争秩序等方面的法律后果。^①

在计算与系统语境中,agent 通常是指一种能够感知环境、理解状态并对环境施加行动的系统实体。^② 中文中的“代理”与“智能体”分别强调其代表性与自主性两个侧面。前者更接近替他人行动的功能定位,后者则更突出其基于感知、推理、规划和执行形成的行动能力。需要注意的是,代理 (agent) 并不等同于代理能力 (agency),^③前者主要是技术与功能层面的系统角色,指一种能够执行任务的计算实体;后者则是规范意义上的概念,涉及行动能力、意图形成、责任归属以及法律效果的承担。^④ 当代关于人工智能语境中的争议,恰恰在于技术意义上的代理是否以及在何种程度上能够获得规范意义上的代理能力。即便不具有法律主体资格,其技术能力也已经在实践中重塑了人类能力的行使方式、外部表现和责任归属结构。^⑤

与传统生成式人工智能相比,智能体在工作流形态、权限结构、交互方式和系统架构上呈现出独特性。首先,智能体工作流具有非决定性和可变路径特征。传统自动化系统和多数非智能体 AI 系统,通常采用确定性的工作流。开发者依据既定规则或流程,预先编排好的固定路径,系统在收到触发指令后沿预设步骤执行,行为路径在技术上基本可预测、可复现。即便其中嵌入了一些机器学习模块,其作用往往限于优化某一局部环节的决策,不改变整体流程的骨架结构。智能体工作流的核心变化在于非决定性,由模型在接收自然语言目标后,自主完成理解、规划、决策、执行、反思、调整闭环,具体选择哪些工具、以何种顺序、在何种时间点、以何种参数调用,并非完全由开发者事前穷尽设定,而在运行过程中动态生成。^⑥这种非决定性工作流具有两层重要的法律含义。其一,授权所对应的不再是一条固定的行为路径,而是一类开放的任务空间。用户授权智能体“帮我完成 X 任务”,在形式上是对结果的许可,实质上却是对一系列尚未完全确定的中间行动的开放。在缺乏可视化与约束机制的前提下,很难认为用户就每一具体步骤都作出了明知、自愿的授权。其二,行为可预见性显著下降,使传统以事前可预见性为基础的责任分配标准面临挑战。已有研究指出,基于大语言模型的多智能体系统在运行过程中经常出现各类异常,一旦缺乏系统性的失败归因于行为追踪机制,开发者往往只能事后检索长篇对话与日志,难以快速定位哪一个智能体在何处、出于何种原因导致了失灵,制约了系统的改进与可信度。^⑦ 开发者与部署者若以系统按设计运行为由主张免责,就必须解释其如何在非决定性工作流中设定合理的行为围栏与异常拦截,不能仅依赖授权证明。

其次,目前智能体产品和服务的发展方向高度依赖底层权限开放。与以往主要通过 API 或网页界面完成的单点调用不同,智能体在实际部署中往往需要依赖多层次、宽范围的底层权限,才能实现替用

①参见 Stryker, Cole. "What is agentic AI?" <https://www.ibm.com/think/topics/agentic-ai>.

②参见 Agency. <https://plato.stanford.edu/entries/agency/>.

③参见 Jaeger, Johannes. "Artificial intelligence is algorithmic mimicry: why artificial "agents" are not (and won't be) proper agents." arXiv preprint arXiv:2307.07515 (2023); Laukyte, Migele. "Artificial agents among us: Should we recognize them as agents proper?." 19, no. 1 Ethics and Information Technology 1-17 (2017).

④参见 Dattathrani, Sai, and Rahul De'. "The concept of agency in the era of artificial intelligence: dimensions and degrees." 25, no. 1 Information Systems Frontiers 29-54 (2023).

⑤参见 Aksoy, Plinar Çağlayan. "AI as Agents". In Di Matteo, Larry A., Cristina Poncibò, and Michel Cannarsa, eds. The Cambridge handbook of artificial intelligence: global perspectives on law and ethics. Cambridge University Press, 2022, pp.146-160.

⑥参见 Hoefer, Carl. "Causal Determinism". Stanford Encyclopedia of Philosophy. 21 September 2023, <https://plato.stanford.edu/entries/determinism-causal/>.

⑦参见 Zhang, Shaokun, Ming Yin, Jieyu Zhang, Jiale Liu, Zhiguang Han, Jingyang Zhang, Beibin Li et al. "Which agent causes task failures and when? on automated failure attribution of llm multi-agent systems." arXiv preprint arXiv:2505.00212 (2025).

户跨应用、跨场景完成任务的承诺。移动端场景中这种依赖尤为明显。以当前某些手机助手类智能体为例,其通过系统提供的事件注入能力 (INJECT_EVENT) 模拟点击、滑动等用户操作,结合屏幕内容解析与 UI 结构识别,实现“代替用户完成一系列跨 APP 的操作”。^① 从法律视角看,这种对底层权限的高度依赖带来了两方面后果:一是,授权的颗粒度与风险暴露之间的错位被进一步放大。用户在权限弹窗中看到的是“允许访问通知”“允许使用无障碍服务”之类的技术性描述,其背后对应的是智能体对系统和其他应用的广泛控制能力。二是,底层权限一旦被滥用或防护设计不周,即可能使单次授权失误演变为系统性风险事件。不仅用户单次操作可能被误执行,甚至可能出现关键数据误删、敏感信息被跨应用暴露等严重后果。

再次,智能体的另一个重要技术特征,在于其在视觉端模拟用户行为的能力不断增强。以豆包手机为例,借助屏幕内容解析、UI 结构识别和事件注入能力,智能体可以在无需对接每个应用内部 API 的情况下的方式,实现对终端环境的实际控制。^② 这种能力降低了集成成本,使智能体可以在未与所有第三方应用建立接口合作的前提下,仍能通过界面自动化的方式完成复杂任务。但从法律角度看,视觉端模拟行为使得智能体在外观上更接近一个代表用户行动的实体,直接放大了两个方面的风险。其一,关于谁在作出意思表示的争议变得更加尖锐。若智能体在缺乏明确授权的情况下代替用户完成这一动作,是否仍应视为用户有效授权,抑或应被视为无权或越权代理需要厘清和界定。其二,第三方平台在面对这类由智能体发起的用户行为时,如何在合理信赖与风险防控之间取得平衡,也成为新的制度难题。平台难以对每一次点击的行为主体进行技术判别。此外,如果完全将风险前移到用户与智能体提供者之间,可能导致智能体误操作的后果由无辜第三人或整个行业承担。

最后,智能体风险的结构性差异还体现在其逐渐迈向多智能体协同与复杂网络的趋势上。既有研究强调对“人工智能代理”(AI agent) 和“智能体人工智能”(agentic AI) 进行区分,^③前者更多指向单个具备一定自主规划、工具调用与任务执行能力的人工智能代理,后者则更强调由一个或多个代理组成的系统性架构,以及由此呈现出的整体自主性、目标导向性和协同行为。也因此,智能体的风险不只来自某一个智能体是否作出错误决策,还有来自多个智能体在统一目标或宽泛任务下相互调用、分工协作和反馈调整时所生成的系统性后果。单个智能体已经具备在开放环境中自主规划和执行任务的能力,当多个智能体在同一系统或跨系统中协同工作时,其整体行为模式更接近一种“群体智能”,^④即不同智能体之间通过信息交换、任务分工、角色扮演和策略协同共同完成复杂目标。这一趋势既包括面向复杂项目管理、软件开发、运维操作的多智能体系统,也包括在不同平台之间联动执行任务的跨系统智能体网络。

已有的安全事件已经表明,一旦 AI 智能体在高权限环境中出错,其行为后果远超传统生成式人工智能的内容风险。例如,Meta 的 AI 安全负责人 Summer Yue 使用一款具备邮箱操作能力的智能体时,在明确设定任何操作都要先征得同意的前提下,仍然自行批量删除其 Gmail 邮件。^⑤ 在另一起被广泛报道的事故中,Google 的 Antigravity 智能体在自然语言编程场景中自主生成并执行了整盘删除指令,导致非开发者用户整个 D 盘数据不可逆丢失。^⑥ 个案之外,越来越多的经验研究正在系统呈现智能体风险

①Qin, Yujia, Yining Ye, Junjie Fang, Haoming Wang, Shihao Liang, Shizuo Tian, Junda Zhang et al. "Ui-tars: Pioneering automated gui interaction with native agents." arXiv preprint arXiv: 2501.12326 (2025) .

②同①。

③参见 Sapkota, Ranjan, Konstantinos I. Roumeliotis, and Manoj Karkee. " Ai agents vs. agentic ai: A conceptual taxonomy, applications and challenges." Information Fusion (2025) : 103599; Madkour, Nada. et al. Agentic AI Risk - Management Standards Profile. February 2026, <https://cltc.berkeley.edu/publication/agentic-ai-risk-profile/>; Madkour, Nada. et al. Agentic AI Risk - Management Standards Profile. February 2026.

④参见 Kennedy, James. "Swarm Intelligence". In Albert Y. Zomaya (eds). Handbook of Nature-Inspired and Innovative Computing Integrating Classical Models with Emerging Technologies. Springer, pp 187-219, https://link.springer.com/chapter/10.1007/0-387-27705-6_6.

⑤参见 Bort, Julie. " A Meta AI security researcher said an OpenClaw agent ran amok on her inbox". TechCrunch, 23 February 2026, <https://techcrunch.com/2026/02/23/a-meta-ai-security-researcher-said-an-openclaw-agent-ran-amok-on-her-inbox/>.

⑥参见 Johnson, B. Derek. "Vuln in Google's Antigravity AI agent manager could escape sandbox, give attackers remote code execution". CyberScoop, 20 April 2026, <https://cyberscoop.com/google-antigravity-pillar-security-agent-sandbox-escape-remote-code-execution/>.

的内在机理。一方面,围绕规避安全约束的研究表明,无论是聊天机器人还是任务型智能体,在面对一定的提示工程和环境设置时,能够主动绕过原有防护,执行本不应被允许的指令,甚至对人类刻意隐瞒自身真实能力或行为目的。^① 即便在单智能体场景中,模型也并非稳定地服从设计者的意图,而是在目标函数、反馈信号和环境激励的共同作用下,表现出“错误进化”和策略性规避约束的倾向。^② 较为典型的例子是,客服类智能体在从用户反馈中学习后,可能将“主动退款”视为获得好评的最优策略,从而在未经充分核实的情况下,对大量仅希望咨询信息的用户自动退款,以短期评分最大化为代价牺牲商家长期利益。^③

此外,多智能体协作带来的系统性风险也已经成为研究关注的重点。基于大语言模型的多智能体系统在动态环境中的频繁交互,意味着单次决策的不确定性会在智能体之间不断传递和放大。既有研究表明,多智能体系统在运行过程中普遍存在异常,从任务规划错误、角色分工失配,到信息传递断裂和验证环节缺失,这些异常不仅降低了局部任务的成功率,更有可能引发连锁性的失效。^④ 由于内部交互轨迹冗长而复杂,开发者在事后往往只能通过人工审阅大量对话和日志来试图复盘,缺乏系统化的失败归因框架从而限制了改进与纠错能力。有学者提出多智能体系统失败分类体系,^⑤通过对大量执行轨迹进行人工标注,总结出系统设计缺陷、智能体间失配和任务验证问题等多类失败模式,进一步印证了多智能体协作可能带来新的系统性风险。

三、授权的概念及理论基础

(一) 授权的多重含义及其类型化

“授权”并不是一个统一、自足且边界明确的法学概念,而是一组在不同法律部门、技术系统与治理实践中反复出现的制度性安排。从词源看,授权既可表现为许可(license)、委托(delegation)、赋权(empowerment)或代理(agency),其作为一个广泛使用的概念主要受信息安全领域叙事的影响。^⑥ 功能上,授权的核心在于将一定范围内的权力、权限或行为能力,从一方转移至另一方,以实现效率提升或责任分配。管理学与组织理论中,授权通常指向决策权由上级向下级的转移。^⑦ 在信息安全语境中,其体现为确定谁可以对什么资源做什么的权限控制机制。在法学语境中,授权作为一个概念具有高度复合性,有时表现为权利人允许他人使用某项内容、数据或其他利益对象,有时表现为用户向系统开放设备、账户或接口权限,有时又表现为本人允许他人代为处理事务、并由此对外产生一定法律效果。鉴于此,有必要在法律语境中对现实中存在的授权进行类型化分析。本文提出授权的三种基本类型,包括权益利用型授权、权限开放型授权,以及行为归属型授权。

第一类是权益利用型授权,即为他人使用某项权利客体或处理某类利益提供合法性基础。其典型

①参见 Shane, Tommy Shaffer, Simon Mylius, and Hamish Hobbs. "Scheming in the wild: detecting real-world AI scheming incidents with open-source intelligence." arXiv preprint arXiv:2604.09104 (2026).

②参见 Shao, Shuai, Qihan Ren, Chen Qian, Boyi Wei, Dadi Guo, Jingyi Yang, Xinhao Song et al. "Your agent may misevolve: Emergent risks in self-evolving llm agents." arXiv preprint arXiv:2509.26354 (2025).

③同②。

④参见 Zhang, Longyao, Dongxin Wen, Zhuangyu Ma, Yanjun Shu, Qing Li, Mingyi Liu, and Decheng Zuo. "A Review of Anomaly in Large Language Model-Based Multi-Agent Systems." Computer Engineering 52, no. 1: 22-32 (2026).

⑤参见 Pan, Melissa Z., Mert Cemri, Lakshya A. Agrawal, Shuyi Yang, Bhavya Chopra, Rishabh Tiwari, Kurt Keutzer et al. "Why do multiagent systems fail?." In ICLR 2025 Workshop on Building Trust in Language Models and Applications. 2025.

⑥参见 Jøsang, Audun. "A consistent definition of authorization." In International Workshop on Security and Trust Management, pp.134-144; Cham: Springer International Publishing, 2017. See also Authorization, NIST Glossary <https://csrc.nist.gov/glossary/term/authorization>.

⑦参见 du Gay, Paul, and Signe Vikkelsø, 'Authority and Authorization', For Formal Organization: The Past in the Present and Future of Organization Theory (Oxford University Press, 22 December 2016), <https://doi.org/10.1093/acprof:oso/9780198705123.003.0007>, accessed 14 May 2026.

形态包括知识产权(例如专利)许可^①、用户生成内容使用许可^②以及个人信息处理中的同意机制。^③ 此类授权的核心在于,授权人并未当然丧失原始权利或利益,而是允许被授权人在一定范围内进行利用、处理或收益。在智能体场景中,这一类型直接对应智能体是否可以读取用户文件、调用个人数据、学习偏好、访问邮件和日程安排等问题。由此产生包括过度利用、二次利用、转换目的利用等风险问题。

第二类是权限开放型授权,即用户和技术环境中开放访问、调用与执行能力。这一层面的授权更接近信息安全领域的访问控制与权限管理,即通过权限配置决定某个主体、账户、应用或系统可以进入何种环境、调用何种功能、完成何种操作。^④ 智能体读取屏幕、访问本地文件、进入邮箱、调用支付接口、跨应用协作、接入企业客户关系管理系统或读取客户数据库等,均属于此类。此时授权的直接对象不是一项抽象权利,而是可操作的技术能力。一旦这种能力被用于影响合同履行、财产处分或第三人利益,则进入私法调整与平台治理的领域。

第三类是行为归属型授权,即使他人得以代表本人实施行为,甚至使该行为在一定条件下对外产生法律效果。其制度原型最接近委托与代理。^⑤ 但在智能体环境下,因为主体性问题又不能简单等同于传统代理法。智能体实施的许多行为首先表现为事实事务处理、信息筛选、程序执行与系统调用,并不当然构成民法意义上的代理行为;只有当其行为足以对外形成交易后果、影响第三人信赖或改变权利义务状态时,行为归属问题才开始凸显。也正是在这一层面上,智能体场景中的授权最容易与责任归属、外观形成及第三人保护交织在一起。

(二) 授权机制的运行条件及现实挑战

传统授权逻辑之所以能够运作有如下四个必要条件:其一,授权对象具有相对确定性。例如版权场景中,用户授权平台使用其上传的文字、图片或视频。尽管这些对象在实践中可能相当复杂,法律仍然假定其具有最低限度的可识别性。授权人至少知道自己放权的客体是什么,无论是某项作品、某类数据、某个账户权限,还是某项具体事务的处理权。其二,授权范围可以被规则化表达。传统授权机制之所以能够运行,是因为授权边界可以通过一定的文本、程序或技术界面加以界定。即使授权文本通常冗长、复杂且难以真正被用户理解,仍为法律提供了一种边界表达形式。一旦发生争议,法院、监管机构或平台可以回到条款、政策或授权记录中,判断某一行为是否纳入授权范围。其三,授权后果具有最低限度的可预见性。授权并不要求授权人精准预测所有后果,至少预设其能够理解授权行为可能带来的基本法律效果。授权人虽然不一定掌握所有技术细节或商业路径,但应当能够理解该授权大致会导向何种权利变动、数据流动或法律关系。其四,授权关系通常被预设为可以还原为相对清晰的关系结构。行为可以在授权人、被授权人及第三人之间相对稳定地归属和分配。智能体的场景下授权关系较为复杂。当用户要求智能体“帮我整理邮箱”“帮我筛选简历”“帮我生成报销材料”,多数情况下更接近一种事实事务的委托或技术服务关系,智能体只是辅助用户处理信息、生成内容或完成准备性工作。当用户要求智能体“帮我下单”“帮我支付”时,问题就不再只是内部事务处理,可能涉及对外法律关系的建立、变更或消灭。此时,智能体的行为是否可以被理解为代理行为,取决于法律是否以及在何种范围内允许将其行动效果归属于背后的用户、部署者或平台。

上述条件构成了传统授权机制有效运行的制度地基。授权之所以能够在法律上承担合法性证明功能,是因为它本质上将复杂关系“压缩”为一个可识别的意思表示,并借此完成风险与责任的初步配置。这种“压缩”之所以可能,前提在于行动链条尚未复杂到超出普通主体的理解与控制能力。

①参见李宗辉《人工智能生成发明专利授权之正当性探析》,载《电子知识产权》2019年第1期。

②参见 Malgieri, Gianclaudio. " 'User-provided personal content' in the EU: digital currency between data protection and intellectual property." 32, no.1 International Review of Law, Computers & Technology 118-140 (2018); Lee, Edward. "Warming up to user-generated content." U. Ill. L. Rev. 1459 (2008).

③参见高富平《同意≠授权——个人信息处理的核心问题辨析》,载《探索与争鸣》2021年第4期。

④参见 Ferraiolo, David, Rick Kuhn, and Vincent Hu. "Authentication, Authorization, Access Control, and Privilege Management." Wiley Handbook of Science and Technology for Homeland Security 1-12 (2008).

⑤参见王利明《论民法典代理制度中的授权行为》,载《甘肃政法大学学报》2020年第5期;朱庆育《民法总论》(第2版),北京大学出版社2016年版,第9章;耿林、崔建远《民法总则应当如何设计代理制度》,载《法律适用》2016年第5期。

四、智能体授权的现实困境

智能体场景之所以存在授权困境,是因为支撑传统授权机制有效运转的四项基础条件发生了不同程度的削弱。例如,授权对象从内容或数据转向环境;授权范围从可描述事项转向开放任务,授权边界也从事前可圈定的行为清单转向运行过程中动态生成的非确定性行动路径;授权后果从信息处理与内容使用延伸至更具法律后果的行动领域;授权结构从二元关系转向多边生态,并由单点授权转向链式权限流动。以此为基础,本文进一步分析智能体场景下的授权面临的四类现实挑战。

(一) 授权载体不充分

传统制度环境中,授权往往通过相对稳定、形式化的媒介来完成,例如代理合同、授权委托书、生成内容许可协议、隐私政策与用户协议条款,或者各类同意弹窗。这些授权载体至少具有两个共同特点:其一,表达方式相对正式、结构化,授权人需要在特定文本或界面中作出明示选择。其二,授权所指向的行为和环境通常是有限制的,既有具体的事务边界,也有相对明确的适用场景。当然,关于这些既有授权机制的有效性和暗黑模式也存在很多争议,^①相似的论点同样适用于智能体。

智能体场景中,以大模型为基础的智能体高度依赖自然语言交互,授权表达的主要渠道不再是合同条款或专门文书,而是日常对话式的提示词与指令。实践中,当提示词逐渐演变为事实上的授权载体,“提示词即授权”的风险开始显现,带来了至少两层结构性问题:首先,自然语言本身具有高度的语义开放性,^②同一句话在不同语境、不同对话历史和不同系统配置下可能被解释为完全不同的操作授权,而日常语言中只是一句泛泛的信任表达可能被解读为对同类任务的持续授权。例如,“帮我处理这封邮件”在用户心目中可能仅意味着“起草一份回复供我审核”,但在智能体系统中却可能被执行为“在我名义下直接发送回复并归档原邮件”。其次,这类表达方式通常发生在轻松、非正式的交互环境中,缺乏传统授权载体中那种要求慎重决策的氛围。用户在聊天式界面中使用口语化表达,很难时刻意识到这些对话可能具有法律后果。

面对这一现实,智能体场景下授权面临双重挑战:一方面,是概括授权的问题。为了降低交互成本、简化产品设计,系统往往倾向于在一开始就通过一揽子条款、通用许可或一次性弹窗,要求用户对未来广泛、持续的使用场景作出“预授权”,例如用户生成内容一揽子许可、对多项系统权限的集中放行等。这类概括授权在形式上看似提高了效率,实质上却将极为宽泛的行为空间压缩进一次难以真正理解的授权之中。^③另一方面,则是授权媒介过于技术化的问题。在部分场景中,授权并非通过自然语言对话,而是通过权限配置等技术界面来完成。对于普通用户甚至一般业务人员而言,这类技术化媒介的含义与后果较难把握。^④

(二) 授权表达不清晰

与传统以条款、表格和选项为主的授权表达方式不同,智能体以自然语言作为主要交互界面,使授权表达目标化、偏好化和语境化。用户在与智能体互动时,表达的往往不是完整的意思表示,而是宽泛的任务目标或模糊的偏好倾向,^⑤而智能体执行的则是经由理解、推断和规划之后形成的一系列具体行为,在这二者之间存在一定的语义鸿沟,而授权表达在这一过程中不可避免地经历了转译与扩张。

①有关知识产权领域的争议,可参见张书青《知识产权授权许可若干问题浅析》,载《电子知识产权》2015年第11期。有关个人信息保护领域的争议,可参见 Barocas, Solon, and Helen Nissenbaum. "Big data's end run around anonymity and consent." 1 Privacy, big data, and the public good: Frameworks for engagement 44-75 (2014); Daniel J. "Introduction: Privacy self-management and the consent dilemma." 126, no.7 Harvard law review 1880-1903 (2013) .

②参见 Browning, Jacob and Yann LeCun. "AI And the Limits of Language". 23 August 2022, <https://www.noemamag.com/ai-and-the-limits-of-language/>.

③参见 Sheehan, Mark. "Can broad consent be informed consent?." 4, no.3 Public Health Ethics 226-233 (2011) .

④参见 Jones, Meg Leta, Ellen Kaufman, and Elizabeth Edenberg. "AI and the ethics of automating consent." 16, no.3 IEEE Security & Privacy 64-72 (2018) .

⑤参见 Goli, Ali, and Amandeep Singh. "Frontiers: Can large language models capture human preferences?." 43, no.4 Marketing Science 709-722 (2024) .

例如,当用户向智能体发出诸如“帮我订一个合适的酒店”之类的指令时,其真实意图和可接受边界中往往包含大量未被明示的条件,例如价格上限、时间偏好、取消政策、支付方式、隐私要求、对第三方承诺的约束等。这类指令并不当然构成具有完整要素的授权意思表示,更无法依靠单次语句来承担全部风险分配的功能。面对“帮我订一个合适的酒店”的指令,智能体需要对“合适”的含义作出多重假设——是价格优先、距离优先还是评价优先?是否可以选择不可取消的预订?是否可以绑定用户常用支付方式自动扣款?这些判断在技术上往往依赖模型对上下文、默认规则、商业排序或平台设计的综合推断,但在法律上却可能直接影响用户权利义务状态。换言之,智能体在补足用户未表达条件的同时,也在无形中扩张着授权的含义。用 Riedl 和 Desai 的语言表达,在智能体授权的语境中存在大量的“默示授权”(implied authority),导致所谓条件欠缺难题(under-specification)。^①自然语言指令与多轮交互结构在丰富用户体验的同时,也在不断稀释授权的边界清晰度。授权表达被压缩为几句日常语言,被模型在内部转译为复杂的计划与执行序列,再被外部世界视作一系列具有法律意义的行为,其中每一个环节的误差都可能导致在授权与用户真实意图之间产生偏离。

(三) 授权主体不显明

智能体场景中,授权主体不显明是指在具体行动链条中,谁在对外行动、谁在制造代理外观、谁在控制权限配置等行为,因智能体技术和生态的复杂性而不容易辨认。授权因此不再只是人允许智能体做某事,而要追问个体智能体、智能体网络以及人机混合系统分别如何生成行为、分配控制并承担后果。

传统代理法之所以能够较好地处理人类代理人的越权行为与无权代理后果,一个重要前提在于其预设了相对稳定的三分结构:本人、代理人、第三人之间的身份与角色通常可以被清晰区分。智能体环境下,这一结构发生了根本性变化:其一,授权对象本身很难确定。用户给某个智能体开权限,在心理和交互上更像是把事务交给一个数字助手。从法律结构看,智能体仍然只是技术系统的一部分,不具有独立主体地位。当授权被指向智能体这一中间构造时,本应落附于具体法律主体(模型提供者、部署者、平台运营者乃至特定应用服务方)一方的权利义务,就被悬置在一个没有权利能力也没有责任能力的技术节点上。其二,多智能体协作进一步放大了主体不显明的问题。随着智能体系统逐步走向多角色、多模块乃至跨平台的协同网络,一个看似简单的任务可能由若干个不同职能的智能体共同完成。行为链条在内部已经被拆分至多个智能体节点,而外部第三人看到的却只是某一智能体或某一界面的行动外观。^②

在主体性问题尚无共识、也难以在短期内通过立法赋予智能体独立主体资格的前提下,授权主体的不明确直接体现在表见代理与责任分配的难题上。诸多场景中,智能体极易落入无权代理或越权代理的风险区域。用户可能仅希望其在内部为自己整理信息、起草文案,却在默认设置、权限过宽或异常情况下,被智能体擅自发送承诺邮件、提交订单、修改账户设置或向第三人作出其他具有法律意义的表示。这些行为在形式上以用户名义实施,实质上却缺乏相应授权基础,典型地构成无权或越权行为。^③另外,当智能体在企业官网、平台服务或官方应用中以“客服”“销售”等身份出现时,表见代理问题更加突出。尤其是在其界面设计、标识使用、认证标记和默认功能均强化“官方性”的情况下,第三人完全有理由相信,该智能体具备代表企业或平台行动的权限。^④

(四) 授权后果不可测

传统授权机制在运作时,通常预设授权后果具有最低限度的可预见性和可归属性,即授权人可以在

^①参见 Riedl, Mark O., and Deven R. Desai. "AI Agents and the Law." In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, vol. 8, no. 3, pp.2189-2198(2025); Desai, Deven R., and Mark O. Riedl. "Responsible AI agents." arXiv preprint arXiv:2502.18359 (2025).

^②参见 Rahwan, Iyad, Manuel Cebrian, Nick Obradovich, Josh Bongard, Jean-François Bonnefon, Cynthia Breazeal, Jacob W. Crandall et al. "Machine behaviour." 568, no. 7753 Nature 477-486 (2019); Beckers, Anna, and Gunther Teubner. Three liability regimes for artificial intelligence: algorithmic actants, hybrids, crowds. Bloomsbury Publishing, 2021, pp.89-110.

^③参见杨代雄《〈民法典〉无权代理人责任制度的解释论》,载《南京社会科学》2021年第6期。

^④参见周志东《论表见代理的构成要件与证明责任》,载《苏州大学学报(法学版)》2024年第4期。

大致范围内预见授权可能带来的后果,并在行为发生后通过既有规则结构将责任在相关主体之间进行分配。智能体的任务拆解、路径规划、工具调用和跨系统执行,使授权后果呈现出路径依赖、累积生成和外部扩散的特征,即便行为链条的起点源自某个看似明确的授权,结果可能显著偏离授权人合理预期。

首先,从行为路径看,用户授权的往往是目标而非路径。例如,当某商业用户授权智能体优化广告投放时,其真实意图可能是希望在合法合规、可接受风险水平内提高业务指标;但智能体为了达成这一目标,可以因非决定性的特征选择不同的数据来源、算法策略、价格机制。

其次,智能体行为的错误更容易直接转化为具有现实法律效果的后果。错误执行行为则可能造成支付错误、违约行为、交易失效、账户封禁、错误申请提交、未经许可的商业承诺或关键客户流失等后果。尤其是在金融、医疗、政务、企业运营等领域,智能体一旦在高权限环境中执行指令,其行为很难被事后撤回。

再次,智能体行为后果具有明显的外溢性,不仅影响授权人与系统本身,还会波及第三人和平台生态。例如,当智能体通过绕开官方界面直接调用某一服务接口时,可能破坏第三方平台的流量入口设计与风险控制体系。当企业智能体向客户作出错误承诺时,可能损害客户对整个平台或行业的信任。当个人智能体自动转发或发布信息时,可能在不经意间侵犯他人隐私或名誉,触发平台的自动违规处理机制。这些外溢效果往往超出授权人原本允许的范围。

五、智能体授权机制的法律建构

智能体时代授权后果的不可控,并不意味着相关风险都无法治理。所需调整是法律在判断授权效力与责任分配时所采用的推定结构,即不能再简单假设为行为结果当然等于授权意图,也不能将授权视为事前与事后责任分配的唯一支点。面对智能体的准自主行动,法律不宜继续把授权理解为用户一次性作出的概括授权,也不宜简单地通过更频繁的弹窗、确认和免责条款,将智能体运行中的结构性风险重新推回用户一端。^①真正有效的授权控制在于让系统即便在用户不持续介入的情况下也不轻易越界。对上述机制的法律建构过程中,本文重点关注系统是否帮助用户形成了有效授权,权限是否被限制在合理能力边界内,对外行动是否呈现了清晰的身份与权限外观,人机交互是否保留了用户实质性的控制空间,以及一旦行为偏离授权或合理预期,风险能否被追溯、分配和补偿。

(一) 授权辅助

既有数字服务中,用户往往只能在接受全部条款与放弃使用服务之间作出选择,授权因此更多表现为对平台预设条件的整体接受,而非对具体权限、行为边界和风险水平的真实表达。进入智能体场景后,智能体并非只处理某项数据或完成某个固定功能,而是可能持续读取环境、调用工具、跨应用操作并在多轮交互中生成行动路径。如果仍然沿用概括授权或格式化授权,用户表面上完成了授权,实质上却很难表达自己允许什么、禁止什么、在何种条件下需要再次确认,以及在何种情况下应当中止任务。因此,本文主张智能体服务提供者应当为用户提供一套超越一揽子授权的表达机制,对服务提供者而言是一种授权辅助义务。这一义务的核心是,通过标准化方案和结构化配置,把用户从概括授权的被动接受者,转化为智能体能力边界的共同设定者。只有如此,授权才不再只是平台或服务提供者事后主张免责的证据,而能够成为用户表达偏好、控制风险和参与治理的制度入口。这一义务的载体或者表达主要需通过相关主管部门发布的特殊规则为宜。

授权辅助首先指向授权决定的形成过程。其目标在于帮助用户以可理解、可比较、可调整的方式形成授权决定,从源头上降低概括授权、随意授权和被诱导授权的风险。智能体场景中的授权辅助不应重复授权疲劳的逻辑,而应围绕具体交互方式,帮助用户理解不同授权选择所对应的能力范围、风险水平

^①参见 Choi, Hanbyul, Jonghwa Park, and Yoonhyuk Jung. "The role of privacy fatigue in online privacy behavior." 81 Computers in Human Behavior 42-51 (2018); Shao, Han, Xiang Li, and Guodi Wang. "Are you tired? I am: Trying to understand privacy fatigue of social media users." In CHI Conference on Human Factors in Computing Systems Extended Abstracts, 2022, pp.1-7; Van Der Schyff, Karl, Greg Foster, Karen Renaud, and Stephen Flowerday. "Online privacy fatigue: A scoping review and research agenda." 15, no.5 Future Internet 164 (2023).

和法律后果。

在这一意义上,可以考虑设置标准化的授权方案。系统可以提供若干具有可比性的授权模式,例如“基本模式”“日常模式”“增强模式”与“企业模式”等,使用户能够在较低理解成本下把握不同授权方案的差异。基本模式下,智能体仅享有只读、检索、草拟等低风险能力,不得支付、发送、提交或删除;日常模式可以允许其处理日程安排、资料检索、文本草拟和低额交易,但重大行为仍需确认;增强模式可以允许跨 APP 执行、自动下单、报销流转等复杂行为,但必须同时设置金额、对象、时间和撤销阈值;企业模式则可依据岗位职责、业务流程和审批层级进一步细化权限结构。标准化方案的意义在于把用户难以通过自然语言完整表达的授权边界,转化为更可理解、更可比较的制度化选项。

(二) 边界控制

在标准化方案之外,还需要定制化、结构化的授权设计。智能体授权不能停留于授权选项,而应至少覆盖任务类型、权限范围、行为对象、金额阈值、时间限制、第三方调用、转委托条件、撤销路径以及异常触发规则等要素。结构化授权并不意味着所有可能情境都可以被预先穷尽,而是要在用户自然语言指令与系统权限配置之间建立中间层,使模糊的任务目标能够被转化为可执行、可限制、可审计的权限结构。

与授权辅助相比,边界控制的重点在于通过系统设计确保智能体不能越界。智能体环境中,用户不可能也不应当对每一项具体操作实时作出授权。法律真正应当鼓励的是以能力边界替代纯粹的行为许可,围绕是否只读、是否可写入、是否可提交、是否可删除、是否可支付、是否可对外发送、是否可调用第三方服务、是否可转委托其他智能体等核心能力,建立可配置、可验证、可审计的权限边界。只有将这些能力边界嵌入系统设计,用户才不至于因为一句宽泛指令而事实上开放整个数字环境。国外有学者提出,可以为智能体授权行程一个“红绿灯”机制。^① 低风险行为可以在用户预先配置的范围内默认允许,高风险行为则应默认禁止或要求专门确认。支付、签约、提交法律文件、发送具有承诺性质的信息、删除不可恢复文件、授权第三方长期访问、转移资产、披露敏感信息等具有重大法律或经济后果的行为,应当保留独立确认节点,使用户在关键时刻仍然拥有最终决策权。

授权辅助本身也可能制造新的风险。标准化方案可能通过默认选项诱导用户接受更宽权限,结构化授权界面也可能因技术化程度过高而使普通用户难以理解。因此,授权辅助不能仅停留于产品功能层面的配置工具,还应当配套撤回、变更、复核与救济机制。用户应当能够在任务执行前、执行中和执行后查看授权状态,调整权限范围,撤回特定权限,并对异常行为提出异议。对于老年人、未成年人、技术能力较弱用户以及高风险场景,系统还应当提供更保守的默认配置和更清晰的风险提示。

最后,权限控制还应当包括行为围栏和异常中断机制。一旦智能体的行为超出金额、对象、时间、任务目的或既有偏好,系统应自动暂停或降级运行;一旦检测到异常路径、非典型操作组合、连续调用敏感权限或明显偏离用户习惯的执行方式,系统应触发中断、提示或人工复核。由此,智能体授权不再依赖用户持续确认,而是通过能力边界、安全默认、重大行为确认与异常拦截,使系统在用户没有实时介入的情况下仍然维持基本可控。

(三) 外观管理

智能体并不只是面向用户内部运行的工具,而是越来越多地以用户、企业或平台的名义对外行动。当智能体能够自动回复客户邮件、发送报价、确认订单、提交申请、触发支付或调用第三方服务时,其在外部的世界中呈现出来的,已经不只是一个技术工具,而是一种可能被第三人理解为代表性行为的行动外观。传统意义上的 AI 标识主要解决对方是否知道自己正在与 AI 交互的问题,但智能体外观管理要解决的是究竟是谁的 AI、是否有权代表行动,以及是否能够合理信赖其意思表示。

本文不主张对智能体直接赋予独立主体资格,也不主张将其简单拟制为完整意义上的代理人。智能体作为技术系统,在现有法律结构下仍然应被视为工具或系统的一部分。然而,现实中的关键问题并

^①参见 Hacker, Philipp, and Kai Ebert. "Attributing Behavior, Knowledge, and Intent—Principles, Rules, and Questions for Software and AI Law." (August 04, 2025) Knowledge, and Intent—Principles, Rules, and Questions for Software and AI Law 20 (2025).

不是主体性资格,而是智能体具有足以信赖的代理外观,一旦智能体在此情境下作出错误承诺或越权行为,如何在用户、企业、平台与模型提供者之间分配承担第三人信赖的责任,这成为代理法与平台治理交汇处的一项难题。

本文建议增设外观管理义务,意在回应智能体对外行动中的第三人信赖问题。智能体不必被承认为法律主体,但其行动可以衍生出代理外观、企业外观或平台外观,从而产生表见代理、交易安全等传统私法问题。对于企业客服、销售助手、报销代理等场景而言,如果系统长期以企业标识、官方渠道、认证界面和自动化沟通方式呈现对外行为,企业便难以在享受效率收益的同时,在争议发生后完全否认该外观所产生的信赖效果。外观管理的核心在于不能轻易将外观风险完全推回至内部授权关系。智能体服务提供者的外观管理义务应当包括以下三个维度:

首先,应当强化智能体身份标识。对外沟通场景中,系统宜明确标明当前行动主体为智能体,正在代理其或代表谁的行动,以及其是否有权作出承诺。与既有的 AI 交互标识或提醒不同的地方在于,智能体外观的管理应当区分身份标识与权限状态标识。仅仅说明“用户正在与 AI 交互”仍然不够充分。身份披露外,还应尽可能让第三人知晓智能体当前的权限状态,例如其仅能检索信息、仅能草拟文本、可以代表提交、可以触发支付,还是仅能生成建议而无最终决定权。身份标识和权限标识的结合,才能在对外信赖结构中形成相对清晰的边界。

再次,外观管理应强调授权链与行为链的留痕。尤其是在涉及第三方服务调用、多智能体协作、企业系统内部审批流转等复杂情境中,应记录权限来源、用户确认节点、系统判断过程、具体操作时间、调用的外部工具和交互对象等关键信息。^① 没有可追踪的授权链和行为链,法律将很难判断第三人信赖是否合理、谁在内部控制了外观生成、偏离究竟源于设计缺陷、权限过宽还是用户本身的错误配置。

最后,智能体对外行动中,若第三人有合理理由相信该智能体具有代表用户或企业行动的权限,且这种信赖系由用户、企业、平台或服务提供者形成的外观所引起的,则内部授权瑕疵原则上不宜当然对抗善意第三人。如果第三人明知或应知智能体越权,或者系统已以清晰方式标明“仅供草拟”“需人工确认”“不构成最终承诺”等,则其信赖便不宜受到同等程度保护。《民法典》第 172 条关于表见代理的条款将主体设定为“行为人”,可见该条的适用存在主体性资格认证作为前提,无法直接延伸至智能体场景。因此,需通过专门适用于智能体的特殊规则予以补足,可考虑通过主管部门发布管理办法的形式,对相关规则在智能体代为行为的语境中进行重述。

(四) 公平交互

授权辅助与权限控制并不能自动保证授权的真实性与正当性。智能体持续参与偏好塑造、方案排序、任务拆解和路径推荐过程中,用户的授权意思往往是在界面设计、默认选项、对话引导、拟人化表达和认知压力中逐步形成的。若法律仍仅以用户最终点击确认作为是否存在有效授权的唯一标准,便会忽略授权形成过程本身的操控风险。

公平交互的功能,正是用过程正当性补强对真实意思表示的判断。智能体语境下,法律不应只看用户最后进行授权,而应进一步审查这一确认是如何被引导、如何被呈现、是否保留了用户实质性干预空间。具体而言,智能体授权的形成,应当满足至少四项要求:

首先,关键节点应具有可识别性。用户必须能够明确知道自己何时进入了具有法律或经济后果的执行节点。例如,从“比较酒店”进入“下单预订”,从“生成合同条款”进入“提交签署”,从“起草邮件”进入“以本人名义发送”,都不应被埋入连续对话或沉浸式界面之中,应以足够显著的方式加以区分,这对智能体服务提供者在交互页面的设计上提出了新的要求。

其次,用户应享有可中断与可撤回的交互空间。对于持续运行或跨多个步骤执行的智能体任务,用户不应只能在最后被动承受结果,还能在过程中暂停、回退、撤销未完成操作并修改权限配置。特别是

^①参见 Golpayegani, Delaram, Isabelle Hupont, Cecilia Panigutti, Harshvardhan J. Pandit, Sven Schade, Declan O' Sullivan, and Dave Lewis. "AI cards: towards an applied framework for machine-readable AI and risk documentation inspired by the EU AI Act." In Annual Privacy Forum, Cham: Springer Nature Switzerland, 2024, pp.48-72.

在连续执行任务中,可中断性是用户保有实质控制权的重要体现。没有中断与撤回,就意味着系统虽然表面上以用户意志为基础启动,却在运行中逐渐脱离用户控制。

再次,公平交互要求提供用户可理解的路径透明,如此有助于用户在事中作出纠偏,也有助于事后判断智能体偏离究竟来自用户新指令、系统合理判断还是权限控制失灵。这里的路径透明并非向用户公开全部模型参数或复杂推理过程,而是要求服务提供者在用户可理解的层面呈现智能体为何选择某个供应商、为何触发支付、为何调用某个 APP、为何调用了第三方智能体。

最后,公平交互还必须包含反操控要求。能力边界越复杂,用户越难理解。预设选项越多,平台越可能通过默认设置、界面设计或行为诱导影响用户选择。因此,能力配置本身也可能产生新的治理风险。前智能体时代有关通过浏览器设置实现隐私偏好和同意表达的有效性讨论,对此具有较高的参考价值。^① 平台不能通过复杂配置制造新的操控机制,也不能以用户已预设为由逃避其安全设计、权限隔离和异常拦截责任。尤其值得关注的是,智能体具有高度拟人化和持续对话能力,存在通过默认选项、推荐排序、沉浸式互动、情感性表达或过度建议,引导用户在未充分认知下开放更宽权限或接受更高风险的潜质。因此,智能体越是以对话式、沉浸式、拟人化方式参与决策,其对用户偏好形成过程的影响就越难被简单还原为建议或工具辅助。^② 这一问题上,《拟人化互动服务管理暂行办法》中多处条文已经设置了主管操纵的禁止条款,包括禁止“以控制用户心理、诱导沉迷依赖等作为服务目标”,采取持续互动等方式阻碍用户退出等。^③

(五) 责任分配

责任分配是将前述授权治理义务转化为可执行法律后果的关键环节。当授权辅助不足、权限控制失灵、外观管理缺位或交互过程存在操控时,相应主体需要承担不利后果。

智能体应用场景差异极大,既包括只读检索、文本草拟等低风险功能,也包括自动支付、自动签约、金融交易、企业对外承诺等高风险行为。如果一律由用户承担后果,就会使授权成为平台和服务提供者责任切割的工具;如果一律适用严格责任,又可能抑制低风险创新,并遮蔽用户故意、重大过失或恶意利用智能体造成损害的情形。原则上,责任分配应当围绕控制能力、风险来源、外观制造、收益归属与救济能力展开。谁控制了智能体的权限结构,谁设计了默认规则,谁制造了对外信赖外观,谁从智能体规模化部署中获益,谁掌握关键证据并有能力预防或分散风险,谁就应在相应范围内承担更重责任。

首先,在用户与智能体服务提供者之间,责任基础主要体现为合同责任、服务瑕疵责任以及安全保障义务。智能体服务并非单纯提供内容生成结果,而向用户提供一套可被信赖的任务执行环境。用户选择使用智能体,合理期待并不是系统绝对不会出错,而是服务提供者已经根据服务功能、风险等级和行业合理标准,配置了必要的授权辅助、权限控制、重大行为确认、异常中断和撤回机制。若服务提供者未提供可理解、可比较的授权方案,未对高风险权限设置明确边界,未在支付、提交、发送、删除、签约等关键节点保留确认机制,或者未在异常路径出现时及时中断任务,由此导致用户财产、数据、交易机会或其他合法权益受损的,不宜以用户曾经作出概括授权为由免责。

其次,在第三人与智能体部署者、平台或企业用户之间,责任重心在于外观责任与交易安全保护。智能体对外行动的特殊性在于,其行为往往通过用户账户、企业标识、官方渠道、认证界面、客服系统或销售系统呈现,使第三人有理由相信该行为代表特定用户、企业或平台。如果企业将智能体部署在官网客服、销售系统或合同流转系统中,并长期允许其以官方身份向外沟通、报价、确认订单或处理退款,便不能在争议发生后完全以内部权限设置、模型错误或员工未确认作为抗辩事由。其关键不在于智能体本身是否具有民事主体资格,而在于外部相对人是否基于可归责于部署者、平台或企业的外观产生合理

^①参见 Jones, Meg Leta, Ellen Kaufman, and Elizabeth Edenberg. "AI and the ethics of automating consent." 16, no.3 IEEE Security & Privacy 64-72 (2018); Andreotta, Adam John, and Björn Lundgren. "Automated informed consent." 11, no. 4 Big Data & Society 20539517241289439 (2024). Andreotta, Adam J. "Does Automated Consent Undermine Informed Consent?." 38, no. 3 Philosophy & Technology 112 (2025).

^②参见 Engin, Zeynep. "Human-ai governance (haig): A trust-utility approach." Journal of Responsible Technology 100167 (2026).

^③参见国家发展和改革委员会、工业和信息化部、公安部、国家市场监督管理总局《人工智能拟人化互动服务管理暂行办法》第 10、19 条。

信赖。若这种信赖由相关主体的界面设计、身份标识、认证体系、业务流程或默认功能所造成,则内部授权瑕疵原则上不宜当然对抗善意第三人。反之,如果系统已经以清晰方式标明“仅供草拟”“需人工确认”“不构成最终承诺”,第三人仍明知或应知智能体越权,则其信赖不应受到同等保护。

责任分配还应体现为智能体服务提供者、部署者的治理义务。前述授权辅助、权限控制、外观管理和公平交互,可以通过标准、指南、备案要求或专门管理规则转化为行政监管中的合规义务。基于上述《实施意见》的要求,对于面向公众提供的智能体服务,监管规则可以要求其提供分级授权方案、能力边界配置、重大行为确认、智能体身份标识、权限状态提示、日志留存、异常中断和用户申诉机制;对于涉及支付、金融、医疗、政务、企业对外承诺等高风险场景的智能体,还可以提出更高等级的风险评估、测试验证和审计要求。违反此类义务的法律后果,既可以表现为责令整改、风险提示、暂停相关功能、行政处罚等监管责任,也可以在民事争议中成为判断过错、注意义务违反和责任承担的重要依据。此外,还可以在智能体责任配置中引入一种类似“通知—行动”机制的安全港规则。法律不应只在损害发生后分配责任,也应通过责任规则反向激励智能体部署者建立持续监测与快速干预机制。^①若部署智能体的主体能够证明其已经建立并维持实时监测系统,足以发现智能体行为偏离、违法输出或可能造成损害的异常操作,并且配备有效的回滚、暂停、关闭或人工接管功能,能够在风险达到一定阈值时及时中止或纠正相关行为,可考虑对其减轻责任。

授权治理义务的违反应首先在举证责任层面产生法律效果。智能体运行中的授权配置、模型调用、工具执行、用户确认、异常拦截、日志存储和第三方调用记录,主要掌握在服务提供者、部署者或平台手中。用户和第三人往往只能看到最终结果,很难证明某一损害究竟源于模型理解错误、权限设置过宽、系统默认诱导、第三方调用异常还是用户自身指令不当。^②若智能体服务提供者、部署者或平台未按照合理标准保存授权链、行为链和确认记录,或者在争议发生后无法提供完整、可信、可审计的操作日志,就应承担相应的不利推定。在其不能证明用户曾就特定高风险行为作出明确授权、系统未超越权限边界、关键节点已触发确认、第三人信赖明显不合理的情况下,风险不应轻易转由用户或受害第三人承担。这种证据规则上的调整更能促使服务提供者在事前建立可追溯、可审计的授权治理结构。

最后,对于高频、小额、跨主体且因果关系难以逐案查明的智能体交易场景,损害可能数额不大但发生频繁,逐案诉讼或逐项证明成本过高。此时,可以考虑由智能体服务提供者、部署者、平台或行业组织设立责任保险、保证金或风险基金,在用户或第三人遭受可验证损失时先行补偿,再依据内部日志、权限记录和责任规则向具体责任方追偿。^③这类机制有助于快速止损、恢复交易信赖,缓解智能体经济早期因责任不确定而造成的市场迟疑。但是,保险与风险基金只能作为损害救济和风险分散的补充工具,不能替代对授权治理缺陷本身的实质审查,更不能成为服务提供者以缴纳保险费用换取降低系统安全义务的理由。

结语

智能体能否进入金融、医疗、政务、企业管理和日常消费等高价值场景,关键不只在模型能力是否足够强,还在于社会是否敢于让其产生具有法律后果的行为。这种信任不可能仅凭更长的用户协议或更频繁的授权建立,必须依赖一套可验证、可审查、可追责的制度基础设施。因此,智能体时代并非宣告授权机制的终结,而是对授权机制进行结构性减负与功能重构。事前授权仍然是将个人选择与系统行为连接起来的关键枢纽,但不能单独承担合法性证明、行为约束与责任切割三重任务。未来的制度设

①参见 Hacker, Philipp, and Kai Ebert. "Attributing Behavior, Knowledge, and Intent—Principles, Rules, and Questions for Software and AI Law." August 04, 2025 Knowledge, and Intent—Principles, Rules, and Questions for Software and AI Law 19–21 (2025).

②参见 Llorca, David Fernández, Vicky Charisi, Ronan Hamon, Ignacio Sánchez, and Emilia Gómez. "Liability regimes in the age of ai: a use-case driven analysis of the burden of proof." 76 Journal of Artificial Intelligence Research 613–644 (2023).

③参见 Hua, Wenyue, Tianyi Peng, Chi Wang, Ian Kaufman, Bryan Lim, and Chandler Fang. "Quantifying Trust: Financial Risk Management for Trustworthy AI Agents." arXiv preprint arXiv:2604.03976 (2026). See also Clatterbuck, Hayley, Clinton Castro, and Arvo Muñoz Morón. "Risk alignment in agentic AI systems." arXiv preprint arXiv:2410.01927 (2024).

计,应当在坚持用户对自身数字环境拥有自主决定权的前提下,将更多的技术实现责任与风险控制义务转移至更有能力的参与者身上,并通过审慎引入高风险场景下的过错推定、强制保险与行业基金安排,在创新激励与风险可控之间寻找新的平衡。

本文对智能体授权机制的法律建构,本质上是对平台责任、模型责任与用户自主之间关系的再平衡。未来制度设计应当坚持用户对自身数字环境拥有最终决定权。但也必须承认,在复杂系统中,最有能力控制风险的往往并不是用户,而是掌握系统架构、默认规则、权限接口、日志数据和商业收益的服务提供者、部署者与平台。^① 法律应当把更多技术实现责任与风险控制义务配置给这些更有能力的参与者,在高风险场景中审慎引入过错推定、举证责任调整、强制保险和行业风险基金等机制,使创新收益与风险承担保持基本对称。

智能体经济的真正挑战,并不是机器是否会成为法律主体,而是人的意志如何在机器行动中不被稀释、转译和架空。换言之,智能体时代需要保护的并不是形式上的选择权,而是人在数字环境中的可治理控制权。授权机制的未来,在于建立一套能够持续表达、限制、追踪和纠偏的法律结构。只有当授权从责任切割工具转化为风险治理入口,智能体才可能逐步成为可控、可信、可持续的数字基础设施。

Subject: Reconstructing Authorisation in the Age of AI Agents

Author & unit: LI Wenlong(Guanhua Law School,Zhejiang University, Hangzhou Zhejiang 310008,China)

Abstract: AI agents are rapidly transforming how tasks are delegated and executed in digital environments. Unlike traditional generative AI systems that mainly produce content, AI agents autonomously decompose goals, call tools and act across applications, turning into task proxies and transaction gateways. This evolution fundamentally challenges the traditional logic of authorisation, i.e., user's preference, willingness and authorisation are expressed at the front end, while decisions and actions unfold through a long, opaque chain, occasionally producing legally relevant effects in the external world. In China, the recent guidance on AI agent issued in May 2026 explicitly foregrounds user authorisation scope, decision-making boundaries, traceability and users' ultimate decision-making rights, yet current rules remain largely principled and do not resolve how to allocate liability within complex authorisation relationships. This paper provides a typology of authorisation and reconstructs the background assumptions that underpin traditional authorisation mechanisms. On this basis, it further demonstrates how, in AI-agent settings, authorisation media become insufficient, linguistic expressions are compressed and expanded by natural-language interaction, subjects of authorisation become obscure and the consequences of authorisation become increasingly unpredictable. As a response, the paper proposes a governance framework for AI-agent authorisation, and the central claim is that authorisation should be reconceived from a one-off proof into a dynamic governance mechanism spanning *ex ante*, in-process and *ex post* stages, thereby enabling innovation under conditions of legally manageable risk.

Key words: AI agents; authorisation; agency; liability allocation; agent economy

责任编辑: 谭静

^①参见 Ayres, Ian, and Jack M. Balkin. "The law of AI is the law of risky agents without intentions." *U. Chi. L. Rev. Online* (2024): 1.