法学与社会治理学刊
2026 Vol. 1, No.1，ISSN 3154-746X

智能体“代理”行为的法律归责困境——以 OpenClaw 类 AI 代理人为视角

何静秋

北京科技大学

通讯作者*：何静秋　E-mail：838456783@qq.com

论文信息

关键字

AI 智能体;代理行为；法律归责；OpenClaw；风险分层

摘要

以 OpenClaw 为代表的开源 AI 智能体的快速普及，标志着人工智能从“对话交互”向“自主执行”的范式跃迁。智能体能够自主拆解任务、调用工具、执行多步骤操作，在邮件收发、商业决策、数据处理等场景中发挥着类似人类代理人的功能。然而，现行代理法律制度以“被代理人的意思表示”为核心构造，预设代理行为须以自然人或法人的意志为归依。AI 智能体的自主性、学习性与决策黑箱等特征，使得授权意思表示的认定、超越权限行为的责任归属以及各方主体的责任分配面临根本性挑战。本文通过分析 OpenClaw 类智能体的技术特征，梳理现行代理法在智能体场景中的适用困境，结合杭州互联网法院“AI 幻觉”侵权案等典型司法实践，借鉴欧盟《人工智能法案》与美国各州的立法探索，提出以风险分层理论为基础的责任分配框架。研究认为，应当区分低风险信息处理行为与高风险财务决策、人身影响行为，建立分层级的合规义务体系；引入过错推定责任以缓解受害人的举证困难；在技术层面嵌入可解释性 AI 作为合规要件，构建开发者、部署者与用户之间的连带责任与内部追偿机制。本文旨在为智能体时代的代理法重构提供理论参照与制度建议。

The legal liability dilemma of the "agent" behavior of agents: From the perspective of an OpenClaw AI agent

ARTICLE INFO

Keywords

AI agent Agency behavior; Legal liability; OpenClaw;" Risk stratification

Abstract

The rapid popularization of open-source AI agents represented by OpenClaw marks a paradigm leap in artificial intelligence from "conversational interaction" to "autonomous execution". Agents can independently break down tasks, invoke tools, and perform multi-step operations, playing a role similar to that of human agents in scenarios such as email sending and receiving, business decision-making, and data processing. However, the

Citation：He, J. Q. (2026). 智能体“代理”行为的法律归责困境——以 OpenClaw 类 AI 代理人为视角 [The legal liability dilemma of the “agent” behavior of agents: From the perspective of an OpenClaw AI agent]. *法学与社会治理学刊*, 1(1), 43–56.

https://doi.org/10.64744/law.2026.198.

current agency legal system is constructed with the "expression of the principal's will" at its core, presupposing that agency acts must be based on the will of natural persons or legal persons. The autonomy, learning nature and decision-making black box characteristics of AI agents pose fundamental challenges to the determination of authorized expressions of intent, the attribution of responsibility for acts beyond authority, and the allocation of responsibilities among all parties involved. This article analyzes the technical characteristics of OpenClaw agents, sorts out the application predicaments of the current agency law in the agent scenario, combines typical judicial practices such as the "AI Illusion" infringement case of the Hangzhou Internet Court, and draws on the legislative explorations of the EU's "Artificial Intelligence Act" and various states in the United States, and proposes a responsibility allocation framework based on the risk stratification theory. Research suggests that it is necessary to distinguish between low-risk information processing behaviors and high-risk financial decision-making and personal impact behaviors, and establish a hierarchical compliance obligation system. Introduce the presumption of fault liability to alleviate the difficulty of the victim in providing evidence; Embed explainable AI as a compliance requirement at the technical level, and establish a joint liability and internal recovery mechanism among developers, deployers and users. This article aims to provide theoretical references and institutional suggestions for the reconstruction of agent law in the agent era.

一、引言

（一）研究背景：AI 智能体的快速普及及其行为替代现实

2025 年底至 2026 年初，一款名为 OpenClaw 的开源 AI 智能体工具迅速席卷全球科技圈，其红色龙虾图标深入人心，“养龙虾”成为技术社群的热门语汇。与传统聊天机器人不同，OpenClaw 可通过整合调用通信软件和大语言模型，在用户本地电脑自主执行文件管理、邮件收发、数据处理等复杂任务。英伟达 CEO 黄仁勋将其称为“有史以来最重要软件发布”，其核心突破在于从“会聊天”向“会干活”的范式跃迁——智能体不再是传统的对话式 AI，而是具备自主执行能力的“数字员工”，可读取文件、发送邮件、运行代码、自动完成多步骤任务。

以 OpenClaw 为代表的 AI 智能体，正在深刻改变人类与技术的互动方式。它们不再仅仅是人类的“工具”，而日益成为能够独立行动、自主决策的“数字代理人”。在商业领域，智能体可以自动处理客户退款、优化供应链调度、执行金融交易；在法律服务领域，智能体可以协助合同审查、法规检索、案件分析；在政务与国企场景中，智能体可能被部署用于处理公共服务、数据分析等职能。这种从“对话交互”走向“自主执行”的跨越，使智能体具备了传统软件工具所不具备的“行动力”。

然而，技术突破与法律规制之间的张力也随之浮现。当智能体自主执行行为导致他人损害时，其法律性质如何界定？责任应当由谁承担？是下达指令的用户，提供底层模型的开发者，还是部署智能体的平台？现行法律框架能否妥善回应这些问题？这些疑问构成了本文研究的核心关切。

（二）问题提出：现行代理法律制度的适用困境

传统代理法律制度以“被代理人的意思表示”为核心构造。根据代理法的基本原理，代理关系的成立以被代理人授予代理权为前提，代理人在授权范围内以被代理人名义实施法律行为，其法律效果直接归属于被代理人。代理权的来源——无论是明示授权、默示授权还是表见授权——都要求存在某种可归因于被代理人意志的授权行为。代理法预设的核心图景是：一个具有法律人格的自然人或法人（被代理人），委托另一个具有法律人格的自然人或法人（代理人），在授权范围内实施法律行为。

AI 智能体的出现，对这一预设图景构成了多重冲击。首先，智能体不是法律主体，不具备独立的法律人格，无法成为代理法意义上的“代理人”。有学者明确指出，在现行法律框架下，AI 智能体不被视为独立民事主体，而应被理解为具备自动执行能力的软件工具或自动化系统。但问题在于，智能体的行为模式已远远超出传统“工具”所能涵盖的范围——它能够自主拆解目标、选择执行路径、调用外部资源，甚至可以根据环境反馈调整自身行为。这种“自主性”使得将智能体简单归为“工具”的做法，在理论上和实践上都显得捉襟见肘。

其次，智能体行为的高度不可预测性使得“授权范围”变得模糊不清。传统代理法中，代理权限通常以明确的书面或口头方式加以界定。但 AI 智能体的学习能力和动态适应性意味着，即使用户在初始部署时设置了明确的任务目标和行为约束，智能体也可能通过自主学习、推理和决策，采取超出用户预期的行动路径。这种“动态授权”与“预设授权”之间的张力，使授权边界的认定变得极为困难。

第三，智能体的“黑箱”特征使得归责过程面临事实认定的障碍。AI 智能体，尤其是基于大语言模型的智能体，其决策过程往往难以被外部观察者所理解和追溯。当损害发生时，究竟应当归咎于用户的指令、模型的设计缺陷、训练数据的偏差，还是智能体在运行过程中的“突发”行为？这种归责链条上的不确定性，构成了法律适用的根本性障碍。

（三）核心问题：智能体代理行为的法律性质与责任分配

基于上述分析，本文的核心问题可以归纳为两个相互关联的层面。第一，AI 智能体的代理行为在法律上应当如何定性？智能体在自主执行过程中所形成的“法律行为”（如发送邮件、订立合同、处理数据），其法律效力能否及于用户？如果能够，其依据何在？如果不能，用户是否需要承担替代责任？第二，当智能体行为致损时，责任应当如何在用户、开发者、平台等多元主体之间进行分配？是否存在一种既能保障受害人获得救济，又能激励技术创新和风险防范的归责机制？

这两个问题的回答，既涉及代理法基础理论的重新审视，也关乎侵权责任法、产品责任法以及数据安全法等相邻法律领域的协调适用。本文试图在梳理技术特征与法律困境的基础上，提出一个兼具理论合理性与实践可行性的责任分配框架。

(四)研究方法与结构安排

本文采用规范分析与比较法研究相结合的方法。在规范分析层面，本文以代理法的基本原理为分析工具，考察智能体行为与代理制度核心要素之间的兼容性；同时借助侵权责任理论，探讨多元主体间的归责逻辑。在比较法层面，本文考察欧盟《人工智能法案》及相关指令草案、美国各州的立法探索以及中国司法机关在相关案件中的裁判实践，为国内法律制度的完善提供参照。

本文的结构安排如下：第二部分阐述 AI 智能体的技术特征与代理行为的类型化；第三部分分析现行代理法在智能体场景中的归责困境；第四部分评析国内外相关司法实践；第五部分提出责任分配框架的重构路径；最后为结论。

二、AI 智能体的技术特征与代理行为类型化

（一）智能体的自主性、学习性与“黑箱”问题

要理解智能体行为对代理法的挑战，首先需要把握其区别于传统软件的核心技术特征。开源智能体是指基于开源协议开放核心代码，以大语言模型作为推理核心，具备自主感知、目标拆解、工具调用、闭环执行能力的人工智能系统。相较于传统生成式人工智能只能进行信息输出，开源智能体可以实现跨时序的状态感知与经验调用，直接完成任务并交付。

自主性是智能体最核心的技术特征。智能体能够将复杂指令拆解为多个步骤，然后调用本地系统资源、外部 API 与第三方插件，完成文件管理、数据处理等复杂任务的闭环执行。这种自主性意味着，智能体在执行过程中可能选择用户未曾预期甚至未曾授权的行动路径。例如，一个被授权“预订符合预算的机票”的智能体，可能在多个订票平台之间自动比较价格、发起交易、完成支付，而这些子操作的具体内容、对象和方式，在用户下达初始指令时往往无法精确预见。

学习性则进一步放大了行为的不确定性。智能体能够通过记忆和规划模块，从历史交互中总结经验，并在后续任务中优化行为策略。这意味着同一个智能体在不同时间点面对相同情境，可能采取截然不同的行动。这种非确定性行为（non-deterministic behavior）对法律的可预见性原则构成了直接挑战——如果行为结果无法被合理预见，过错和因果关系等归责要件将难以成立。

“黑箱”问题则是前述特征叠加后的必然产物。AI 智能体，特别是基于深度学习模型构建的智能体，其决策过程往往具有高度的复杂性和不可解释性。正如有学者指出，AI 系统因其自主学习的特性，能够发现人类观察者难以察觉的数据间隐藏关联，这使得理解系统如何或为何得出特定结论变得极为困难。这种不可解释性，在司法证明层面构成了受害人和法院的认知障碍——谁有权要求智能体的开发者和部署者披露决策过程？如何确保这种披露在技术上可行且在法律上可强制执行？

（二）典型代理场景：邮件收发、商业决策、数据操作、合同草拟

AI 智能体在现实中的应用场景广泛而多元，其中以下四类与代理法律关系的关联尤为密切。

邮件收发是 OpenClaw 类智能体最为基础也最为普遍的应用场景。智能体可被授权在用户本地电脑上自主收发邮件、管理收件箱、根据指令内容自动回复或转发邮件。在这一场景中，智能体的行为直接涉及用户与他人的通信往来，可能产生法律上的意思表示效果。例如，当智能体自动回复一封包含承诺内容的工作邮件时，该回复是否应被视为用户本人的意思表示？这直接触及代理法中的授权与表见代理问题。

商业决策是更具风险性的应用场景。智能体可被授权执行客户退款处理、供应链优化、库存管理、投资组合调整等商业操作。在这些场景中，智能体的决策可能直接影响企业的财务状况和与第三方的法律关系。例如，一个被授权管理公司账户的 AI 智能体，若因错误解读市场信息而进行了一笔失败的投资，导致公司巨额亏损，责任应如何划分？

数据操作是另一个高频应用场景。智能体可被授权读取、处理、分析甚至跨境传输用户的各类数据，包括个人数据、商业秘密和敏感信息。这种授权往往以“默认全量授权”或“持续读取全部历史数据”的方式实现，在数据保护法下面临“最小必要”原则的挑战。

合同草拟则涉及法律服务的专门领域。智能体可基于用户提供的案件事实和合同模板，自动生成法律文书、合同条款或法律意见。但问题在于，智能体生成的合同内容是否构成有效的法律文件？谁应为合同内容的准确性负责？如果智能体生成的合同条款存在重大瑕疵导致用户损失，责任归属于谁？

（三）“代理授权”的边界模糊：预设授权与动态授权的冲突

上述应用场景共同揭示了一个根本性问题：智能体场景中的“授权”已不再是代理法传统意义上的静态授权概念。传统代理法中，代理权限以被代理人的意思表示为依据，通常在代理关系建立时即已明确界定。代理人的行为是否在授权范围内，是一个可以事后根据授

权文件进行判断的事实问题。

然而，智能体的动态学习和自主决策能力使得“授权范围”成为流动的、不确定的概念。用户可能在初始部署时给予智能体较为宽泛的指令（如“处理客户退款”），智能体在执行过程中通过自主学习不断调整其行为策略，其实际采取的行动可能显著偏离用户的初始意图。这种“预设授权”与“动态授权”之间的冲突，在以下三个维度上尤为突出。

第一，授权的“时间维度”被拉长。传统代理授权是一次性的、事前完成的动作；而智能体授权则是一个持续的过程，智能体在执行过程中不断“获得”新的信息、不断“调整”其行为，其行为的边界不断被重新定义。用户很难在授权之初就对智能体的所有潜在行为进行完整的预判和规范。

第二，授权的“内容维度”变得模糊。传统代理授权通常以“行为类型”或“交易对象”作为界定标准；而智能体授权则可能涉及大量的子操作和中间步骤，这些子操作的性质和风险程度可能与用户的初始指令存在显著差异。用户可能授权智能体“处理客户信息”，却未曾预料到智能体会将数据跨境传输给第三方服务商。

第三，授权的“主体维度”发生漂移。在多智能体协同部署或智能体与第三方插件交互的场景中，原始授权可能在智能体与智能体之间传递和转化，导致最终执行行为的智能体与最初授权的用户之间形成了多个“代理层级”。这种链条的延长，使得授权链条的完整性难以保证，也使得责任归属变得异常复杂。

三、现行代理法律制度面临的归责困境

（一）授权意思表示的认定难题：用户点击“同意”是否构成有效授权

代理法的起点是“授权”。但智能体场景中，用户的授权行为应当如何认定？一个常见但值得深究的问题是：用户在注册或使用 AI 智能体服务时点击“同意”用户协议，是否构成代理法意义上的有效授权？

从形式上看，用户点击“同意”确实是一种意思表示，表达了用户接受服务条款、允许智能体在特定条件下执行特定操作的意愿。然而，代理法要求授权意思表示具有特定性和明确性——被代理人应当能够合理预见代理行为的范围和后果。在智能体场景中，用户协议往往采用概括性、格式化的表述（如“用户同意授权本应用程序访问您的设备信息、文件系统和网络资源”），用户对智能体可能采取的具体行动——尤其是那些超出普通用户认知范围的行动——往往缺乏足够的预见。

更根本的问题在于，智能体的自主学习能力意味着，即使协议条款本身是明确的，智能体在运行过程中也可能通过学习不断拓展其行为边界，从而超越协议所预设的授权范围。在这种情形下，用户的点击“同意”是否仍应被视为对该等行为的有效授权？如果答案为否定，那么用户是否应因未能有效监督和约束智能体而承担某种责任？

（二）超越授权的责任归属：用户、开发者、平台如何担责？

当智能体的行为超越用户预设的授权范围而导致损害时，责任的归属问题变得尤为棘手。理论上，潜在的责任承担主体至少包括用户（部署者）、开发者、平台服务提供者以及智能体本身（但后者因缺乏法律人格而被排除）。

用户或部署者通常是首当其冲的责任承担者。有学者指出，在我国现行法下，OpenClaw不属于独立民事主体，而应被理解为具备自动执行能力的软件工具；其对外发生的电子操作，真正需要归责的主体仍然是部署单位、账号持有人、权限配置人、流程审批人以及具体服务提供者。换言之，主张“AI 自己做的”不足以形成责任真空。这一立场体现了“谁控制风险、谁承担后果”的基本逻辑，符合代理法中“被代理人对其代理人负责”的一般原则。

然而，将全部责任不加区分地归于用户，在智能体高度自主的场景下可能产生不公平的结果。试想：用户购买了某公司的智能体服务，设置了看似合理的初始权限和约束，但智能体因模型本身的设计缺陷或训练数据的偏差而采取了危险行动。在这种情况下，用户既没有过错，也无法预见损害的发生，为何要承担全部责任？

这就指向了开发者和平台的责任。有分析指出，法院在进行司法裁判时，不太可能孤立地看待某个环节，而更倾向于在开发者、服务提供商、业务部署方和终端用户这四个关键节点之间，根据各自的过错程度、控制力和受益情况来分配责任。这一判断揭示了归责实践中的基本取向：归责应当是“分配性”的，而非“单一性”的。但问题在于，现行法律缺乏关于如何分配这类责任的明确规则，法院不得不依赖一般侵权法的过错责任原则进行个案裁量。

（三）免责条款的滥用风险：以“技术黑箱”或“AI 幻觉”为由逃避责任

归责困境的另一个侧面，是免责条款的潜在滥用风险。开源智能体项目通常在其许可协议中包含“AS-IS”（按原样提供）和“NO WARRANTY”（不提供任何担保）的免责声明。在商事交易中，法院普遍尊重这些条款，认为商业用户在享受开源软件带来的巨大价值的同时，也应承担相应的审查和使用风险。

然而，当智能体的行为超出用户合理预期，且损害后果严重时，这种“按原样提供”的

免责逻辑是否仍然合理？问题的关键在于，传统软件的行为模式是确定性的——给定相同的输入，软件总是产生相同的输出。用户可以基于这种确定性来合理预见风险并采取防范措施。但智能体的非确定性行为意味着，即使用户采取了所有合理措施，损害仍然可能因无法预见的原因而发生。在这种情况下，简单的免责条款是否仍能被视为对用户公平？

此外，“AI 幻觉”问题使免责条款的滥用风险更加突出。智能体可能生成看似合理但实际不准确或虚假的信息，导致用户或第三人遭受损害。如果开发者或平台以“这是 AI 幻觉，不是我们的过错”为由拒绝承担责任，那么受害人将面临无法获得救济的风险。杭州互联网法院在一起 AI 幻觉侵权案中明确否定了将 AI 生成内容视为平台意思表示的尝试，但该案的逻辑是否能够类推适用于智能体自主执行行为，仍有待检验。

（四）比较法视角：欧盟《AI 责任指令》草案与美国的实践分歧

欧盟选择了以风险分级为基础的综合监管路径。2024 年通过的《人工智能法案》（Regulation 2024/1689）是全球首部全面规范人工智能技术的综合性立法，基于风险分级管理理念，确保 AI 技术在欧盟市场的应用尊重基本权利与安全。对于智能体这类自主性 AI 系统，欧盟的规制思路是将其纳入风险分类框架，根据其潜在危害程度施加相应的合规义务。然而，批评者指出，现行 AI 法案并未充分涵盖 AI 智能体的独特特征。CEPS 的一份政策报告警告称，AI 智能体因其高度的自主性、适应性和环境交互能力，可能挑战 AI 法案中的基础性假设，特别是在“提供者”的界定和责任的分配方面。

美国则呈现出“联邦缺位、各州先行”的碎片化格局。加州 AB 316 法案已于 2026 年 1 月 1 日生效，该法案明确规定，被告不得以 AI 系统的自主运行为由对责任主张提出抗辩。这意味着在加州，企业不能简单地主张“AI 自己做的”来规避责任。这一立法选择体现了将“自主性”本身视为应归责因素的思路，与欧盟的风险分类路径形成鲜明对比。

在中国，尽管尚未出台专门针对 AI 智能体的综合性立法，但现行法律体系已在逐步纳入 AI 治理的要求。2025 年修正并自 2026 年 1 月 1 日起施行的《网络安全法》，已将人工智能风险监测评估和安全监管纳入法律文本。这表明，中国立法机关已经意识到 AI 技术带来的监管挑战，并在现有法律框架内寻求回应方案。然而，针对智能体代理行为的专门归责规则，目前仍处于制度空白状态。

四、国内外司法实践与案例评析

（一）杭州互联网法院“AI 幻觉”侵权案（2025）

2025 年底，杭州互联网法院审结了梁某与某科技公司网络侵权责任纠纷一案，该案被

广泛视为“全国首例 AI 幻觉引发的侵权纠纷案”。案件的基本事实是：原告使用被告运营的生成式 AI 应用程序查询某高校报考信息，AI 生成了关于该校主校区的不准确信息。原告纠正后，AI 仍坚持错误信息，并生成了一段内容，表示若生成内容有误将向用户赔偿 10 万元，并建议原告到杭州互联网法院起诉索赔。原告据此起诉，要求被告赔偿 9999 元。

法院的裁判要点包括两个核心层面。第一，法院明确认定，人工智能不具有民事主体资格，不能作出意思表示。法院指出，现行法中享有民事权利、能够作出意思表示的民事主体仅有三类：自然人、法人和非法人组织。人工智能模型既不是生物学意义上的人，也未被我国现行法律赋予民事主体资格，因此不能独立、自主地作出意思表示。基于此，法院认为 AI 生成的“承诺”内容不能成立为 AI 或其服务提供者的意思表示，不产生意思表示的法律效力。

第二，关于侵权责任的适用，法院认为此类生成式 AI 服务引发的侵权纠纷应适用一般侵权责任的过错责任原则，而不应适用产品责任的无过错责任原则。理由在于：案涉生成式 AI 属于服务而非产品，不具备具体、特定用途及合理可行的质检标准，难以适用产品责任。法院最终判决驳回原告的诉讼请求。

这一判决对理解智能体行为的法律定性具有重要参照意义。首先，法院明确拒绝了赋予 AI 以意思表示能力的思路，坚守了“人”才是法律主体的基本立场。这一立场与代理法的核心预设是一致的，即在没有法律人格的情况下，AI 不能成为代理法意义上的“代理人”。其次，法院将生成式 AI 服务定性为“服务”而非“产品”，排除了产品责任的适用，这意味着受害人需要证明服务提供者的过错才能获得赔偿。在智能体的“黑箱”特征下，这一举证负担可能成为受害人获得救济的实质性障碍。

然而，本案的局限在于，其处理的是“信息输出”场景下的 AI 幻觉问题，而非智能体“自主执行”致损的问题。在信息输出场景中，AI 的行为停留在“建议”层面，尚未进入“行动”层面；而在智能体自主执行场景中，AI 的行为直接干预外部世界，造成物理的或经济的损害，其法律后果的严重程度远非信息误导可比。杭州互联网法院的判决逻辑能否延伸适用于自主执行场景，仍有待司法实践的进一步检验。

（二）美国相关判例（如 Smith v. OpenAI，2024）

在大洋彼岸，美国法院也在逐步面对 AI 智能体引发的归责问题。一个值得关注的动向是加州北区联邦法院近期处理的一起集体诉讼案件。在该案中，法院批准了针对某 AI 驱动招聘平台的集体认证申请，原告指控该平台使用的 AI 推荐系统在评分、排序、筛选和筛选

申请人时，存在对年长求职者的系统性歧视。法院认为，原告充分指控了使用 AI 推荐系统作为统一政策的事实，由此认定的集体可能涵盖数亿受影响申请人。

这一案件的重要启示在于：当 AI 系统在具有重大影响的决策场景中运行时，部署该系统的组织将面临司法审查，并可能需要承担相应责任。法院并未接受“算法黑箱”或“AI 自主决策”作为免责理由，而是将 AI 的决策视为组织的“统一政策”的体现。这一立场与加州 AB 316 法案的精神一脉相承——禁止以 AI 的自主运行为由规避责任。

此外，学者们还指出了另一类美国判例的潜在影响：在涉及 AI 系统缺陷导致损害的产品责任诉讼中，美国法院可能对传统产品责任法进行扩张解释，将 AI 系统纳入“产品”范畴。这与欧盟新版产品责任指令将软件纳入产品范围的立法动向形成呼应，但美国联邦层面目前尚无统一立场。

（三）案例启示：归责规则缺失，裁判标准不统一

上述案例揭示了一个共同的困境：在缺乏明确归责规则的情况下，法院被迫依赖一般侵权法原则进行个案裁判，这不仅导致裁判结果的不确定性，也可能阻碍受害人的权利救济。杭州互联网法院的判决虽然澄清了 AI 不具意思表示能力这一重要问题，但并未解决智能体自主执行致损时的归责难题。美国法院虽然倾向于将责任归于部署者，但这一做法是否以及在多大程度上应被中国法院采纳，仍缺乏理论共识和制度支撑。

裁判标准的不统一，一方面源于法律规范的缺位——我国目前尚无针对 AI 智能体的专门立法，现行代理法、侵权法和产品责任法在智能体场景中的适用存在明显的规范断层；另一方面，也反映了学理层面的分歧——关于 AI 智能体的法律地位、授权行为的认定、责任分配的逻辑等基础性问题，法学界尚未形成主流共识。这种“规范缺位”与“理论分歧”的双重困境，构成了智能体归责制度构建的根本障碍。

五、责任分配框架的重构路径

（一）风险分层理论：区分低风险与高风险智能体行为

面对智能体行为的多样性，一种有前景的重构思路是引入风险分层理论。该理论的核心主张是：不应当对所有智能体行为适用同一套归责规则，而应当根据智能体行为所涉风险的层级和性质，设置差异化的法律要求。

具体而言，可以将智能体行为区分为低风险类型与高风险类型。低风险行为主要限于信息处理、文档管理、常规通信等不直接涉及重大人身权益或财产权益的活动。例如，智能体自动整理用户文件、转发非敏感邮件、生成日常办公文档等。对于这类行为，可以适用相

对宽松的法律要求，以鼓励技术创新和效率提升。用户通过点击同意用户协议所构成的概括性授权，在这一层级上可能被视为足够充分。

高风险行为则包括涉及重大财务决策（如投资交易、支付授权）、人身安全影响（如医疗辅助决策、自动驾驶）、敏感数据处理（如生物识别信息、健康数据）以及法律权利变更（如合同订立、授权委托）的活动。对于这类行为，应当施加更为严格的合规要求。具体而言，包括但不限于：高风险行为必须经过用户明确、具体的单独授权，不得以格式条款或概括性授权替代；部署者应当建立实时监控和干预机制（“人类在环”）；智能体的操作日志应当完整保留并可供事后审计。

风险分层的优势在于，它既避免了“一刀切”规制对技术创新的过度抑制，又确保了高风险场景下受害人权益和公共安全的充分保护。在制度设计上，可以借鉴欧盟 AI 法案的风险分级思路，同时结合我国法律体系的特征进行本土化改造。

（二）强制披露义务：决策逻辑的可解释性与关键日志留存

“黑箱”问题是智能体归责中的核心障碍。要解决这一问题，必须在技术标准和法律规范两个层面同步推进。在技术层面，可解释性 AI（XAI）的发展为智能体决策的透明化提供了可能性。在法律层面，则应当建立强制披露义务，要求智能体的开发者和部署者在一定条件下披露决策逻辑，并保留关键操作日志。

具体而言，强制披露义务可以包括以下要素。第一，对于高风险智能体行为，开发者应当在设计和训练阶段嵌入可解释性机制，确保关键决策可以被追溯和理解。第二，部署者（通常是用户或企业）应当建立完整的操作日志系统，记录智能体的每一次关键操作，包括操作时间、操作对象、操作内容以及触发该操作的输入信息。第三，当损害发生后，受害人向法院申请证据保全时，法院有权要求开发者和部署者提供相关决策日志和解释材料，不得以“商业秘密”或“技术黑箱”为由拒绝。

值得指出的是，强制披露义务的建立不能仅依赖法律强制，还需要技术标准的配套支撑。我国可以借鉴 NIST 等机构发布的技术指南，制定适用于智能体的可解释性技术标准，并将其嵌入法律合规要求之中。

（三）归责规则建议

1.过错推定责任：举证负担的合理分配

在现行一般侵权责任的过错责任原则下，受害人需要证明智能体部署者存在过错，方

可获得赔偿。然而，智能体的“黑箱”特征和决策过程的复杂性使得这一证明在大多数情况下几乎不可能完成——受害人无法进入智能体的“大脑”查看其决策过程，更无法证明部署者的过错与损害之间的因果关系。

为了缓解这一问题，可以考虑在智能体致损的场景中引入过错推定责任。其基本逻辑是：当智能体的自主执行行为导致他人损害时，推定部署者存在过错；部署者如欲免责，须证明自己已经尽到了合理注意义务，包括但不限于：采取了行业内公认的安全措施、对智能体进行了充分的测试和审查、建立了有效的监控和干预机制、损害的发生超出了合理可预见的范围。

过错推定责任的引入，在比较法上已有参照。加州 AB 316 法案禁止以 AI 系统的自主运行为由规避责任，实质上体现了类似的立法精神——将举证负担从受害人转移到部署者一方。当然，过错推定的适用范围需要谨慎界定，以避免对技术创新产生过度抑制。可以将其限定于高风险智能体行为，对于低风险行为则仍适用一般过错责任原则。

2.连带责任设计与内部追偿机制

智能体致损的责任链条通常涉及多个主体：提供基础模型的开源社区或企业、部署智能体的用户或企业、提供计算资源的云服务商、开发第三方插件的技能提供者等。在单一责任主体无法完全赔偿受害人损失，或者过错难以明确归属于某一主体的情况下，连带责任制度可以提供一种务实的解决方案。

具体而言，可以设计如下的连带责任框架：当智能体的自主执行行为导致他人损害，且损害结果与智能体行为之间存在因果关系时，用户（部署者）首先承担对外的赔偿责任。用户赔偿后，有权向智能体的开发者或平台服务提供者追偿，但需证明后者的过错（如模型设计缺陷、训练数据偏差、安全措施不足等）与损害之间存在因果关系。开发者和平台服务提供者之间也可以根据各自的过错程度和风险控制能力进行内部责任划分。

这一机制的优势在于，它确保了受害人在第一时间获得赔偿的权利，避免了受害人被卷入复杂的多方法律关系之中。同时，内部追偿机制为用户和开发者之间的责任分配保留了充分的灵活性和个案裁量空间。

（四）技术标准的法律嵌入：可解释性 AI 作为合规要件

法律规则的有效实施离不开技术标准的配套支撑。在智能体归责领域，一个关键的制度设计是将可解释性 AI（Explainable AI, XAI）的技术要求嵌入法律合规体系，使其成为

智能体开发者和部署者的法定义务。

所谓可解释性 AI，是指 AI 系统的决策过程和输出结果可以被人类理解、追溯和解释的技术特性。在法律语境下，可解释性的意义在于：它为归责过程中的事实认定提供了可能性——如果无法解释智能体是如何作出决策的，法院就难以判断过错的存在和因果关系的成立。因此，将可解释性作为合规要件，实际上是在技术层面为法律归责创造前提条件。

具体可以借鉴欧盟 AI 法案中关于高风险 AI 系统“透明度义务”的规定，结合我国实际情况进行本土化改造。对于被归类为高风险智能体行为的系统，其开发者应当提交可解释性评估报告，证明其系统的关键决策可以被有效追溯和解释。未能达到可解释性标准的智能体，不得被用于高风险场景。这一制度设计将技术合规与法律归责有机衔接，使“技术黑箱”不再是逃避责任的借口。

六、结论

以 OpenClaw 为代表的 AI 智能体的快速崛起，正在深刻重塑人类与技术的互动方式。智能体的自主执行能力——从邮件收发到商业决策、从数据处理到合同草拟——使其越来越接近于传统代理法意义上的“代理人”。然而，代理法律制度以被代理人的意思表示为核心构造，以自然人或法人的法律人格为基本前提，难以适应智能体的自主性、学习性和决策黑箱等技术特征。

本文的分析表明，现行代理法在智能体场景中面临着三重困境：授权意思表示的认定难题、超越授权行为的责任归属困境，以及免责条款的滥用风险。国内外司法实践进一步印证了归责规则缺失、裁判标准不统一的现实问题。在比较法层面，欧盟的风险分级路径、美国各州的立法探索以及中国的初步实践，为制度重构提供了有益参照。

在此基础上，本文提出了以风险分层理论为基础的责任分配框架。核心建议包括：区分低风险信息处理行为与高风险财务决策、人身影响行为，建立分层级的合规义务体系；建立强制披露义务，要求开发者和部署者保留关键日志并提供决策解释；在高风险场景中引入过错推定责任，以缓解受害人的举证困难；设计连带责任与内部追偿机制，确保受害人获得有效救济的同时，在用户、开发者与平台之间合理分配责任；将可解释性 AI 作为合规要件嵌入法律规范，实现技术标准与法律规则的有机融合。

AI 智能体的发展才刚刚开始，法律的回应也必须与时俱进。本文所提出的重构路径，旨在为智能体时代的代理法改革提供一个初步的理论框架，但制度的最终完善仍有赖于学

界的持续探讨和司法实践的不断积累。在技术理性与法律理性的交汇处，我们既需要保持对创新的开放态度，也必须坚守对权利保护和风险防控的基本承诺。

参考文献

[1] Torrance, A. W. & Tomlinson, B. (2025). Decentral Intelligence Agency: The Law and Autonomous Artificial Intelligence. Touro Law Review, 40(4).

[2] Kolt, N. (forthcoming). Governing AI Agents. Notre Dame Law Review.

[3] Nannini, L., Smith, A. L., Maggini, M. J., et al. (2026). AI Agents Under EU Law. arXiv preprint, arXiv:2604.04604.

[4] Sakr, J. (2026). When Algorithms Act: Reconstructing Legal Agency in Digital Governance. Journal of Internet Law, 29(5), 1-23.

[5] Chiruvolu, P. (2025). Agentic AI: Greater Capabilities and Enhanced Risks. Practical Law, Article w-047-6759.

[6] 张莹，殷文超. (2026). 关于 OpenClaw 类自治智能体在政务和国企场景应用中的法律风险分析. 德和衡律师事务所.

[7] 李鑫，周俊杰. (2026). 开源智能体法律风险的审视与治理. 四川法治报，2026-04-03.

[8] 梁对. (2026). 律师该不该“养龙虾”？——律师行业使用 OpenClaw 的利弊分析及对策建议. 北京德恒律师事务所.

[9] 赵暄. (2026). OpenClaw 合规困境：从权限到 AI 征税，行业自律该往哪走？ 曼昆律师事务所.

[10] 天元律师事务所. (2026). “龙虾”出笼，AI 干活：自主智能体的法律挑战与合规路径——以 OpenClaw 为视角.

[11] 马民虎，黄道丽. (2026). AI 智能体专门立法与现行法如何互补融合. 信息安全与通信保密.

[12] Centre for European Policy Studies. (2025). Ahead of the Curve: Governing AI Agents Under the EU AI Act.

[13] Employer’s Vicarious Liability for Damage Caused by an AI Worker: Comparative Law Perspective. (2025). Utrecht Law Review, 21(1), 36-48.

[14] How to Count AIs: Individuation and Liability for AI Agents. (2026). International Center for Law & Economics.

[15] Agency reimagined: Liability in an era of agentic AI. (2026). T G Legal.

[16] 杭州互联网法院. (2025). 生成式人工智能平台侵犯信息网络传播权案一审判决书.

[17] 杭州互联网法院. (2025). 梁某与某科技公司网络侵权责任纠纷案判决书.

[18] California Assembly Bill 316 (2025). AI Autonomy Defense Prohibition.