Inherent and emergent liability issues in LLM-based agentic systems: a principal-agent perspective

Garry A. Gabison[1, 2, *], **R. Patrick Xian**[3, 4, *],
[1]Queen Mary University of London, [2]UC Berkeley, [3]Certivize AI, [4]UC San Francisco
[*]Equal contribution
✉: g.gabison@qmul.ac.uk, xrpatrick@gmail.com

Abstract

Agentic systems powered by large language models (LLMs) are becoming progressively more complex and capable. Their increasing agency and expanding deployment settings attract growing attention to effective governance policies, monitoring, and control protocols. Based on the emerging landscape of the agentic market, we analyze potential liability issues arising from the delegated use of LLM agents and their extended systems through a principal-agent perspective. Our analysis complements existing risk-based studies on artificial agency and covers the spectrum of important aspects of the principal-agent relationship and their potential consequences at deployment. Furthermore, we motivate method developments for technical governance along the directions of interpretability and behavior evaluations, reward and conflict management, and the mitigation of misalignment and misconduct through principled engineering of detection and fail-safe mechanisms. By illustrating the outstanding issues in AI liability for LLM-based agentic systems, we aim to inform the system design, auditing, and tracing to enhance transparency and liability attribution.

1 Introduction

AI agents are computer software systems capable of creating context-specific plans in non-deterministic environments (Chan et al., 2023; Krishnan, 2025). AI agents based on LLMs (aka. LLM agents, see Appendix A) exist on a spectrum of autonomy, ranging from simple tool-calling agents to generalist agents capable of planning, sourcing, critiquing, and executing their own workflow (Li, 2025). They primarily adopt an architecture with explicitly defined functioning components[1] (Sumers et al., 2023). LLM-based multiagent systems (MASs) allow agents to interact, collaborate, or compete within shared environments (Fig. 2a). They are designed by combining agents with specialized roles through LLM role-playing (Shanahan et al., 2023; Chen et al., 2024a) or through integration on a software platform. Each agent handles specific sub-

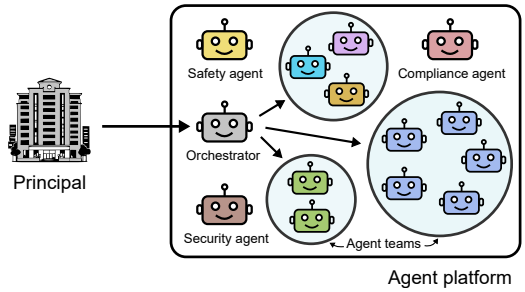

Figure 1: A plausible LLM-based MAS deployed on an agent platform, where delegation goes from the principal to an orchestrator (agent) to different functioning agent teams (circles). The platform also contains supporting agents for safety, security, and compliance. Colors distinguish between agents of different types.

tasks based on its expertise and allocated resources (tools, data, compute, etc). MASs can be tailored to a wide range of scales and domains, from few-agent systems that simulate team decision-making in medicine (Tang et al., 2024; Kim et al., 2024) and finance (Xiao et al., 2025), to many-agent systems that mimic the population-level socioeconomic dynamics (Park et al., 2024a; Piao et al., 2025). A plausible LLM-based MAS (Fig. 2a) can possess multiple teams of interacting agents coordinated by an orchestrator (Wang et al., 2025) and behaviorally regulated by other platform agents or through a set of engineered constraints (e.g. norms) (Criado et al., 2011; Hadfield-Menell et al., 2019). The components mentioned here are further defined and explained in Appendix A. While the flexibility of LLM-based MASs allows adaptation to various applications and affordances, it also introduces emergent risks not present in single agents (Hammond et al., 2025; Pan et al., 2025).

[1]Also called cognitive architecture (Kotseruba and Tsotsos, 2020) at times, but the architecture alone doesn't guarantee cognition or sense of agency.

Proceedings of the 1st Workshop for Research on Agent Language Models (REALM 2025), pages 109–130

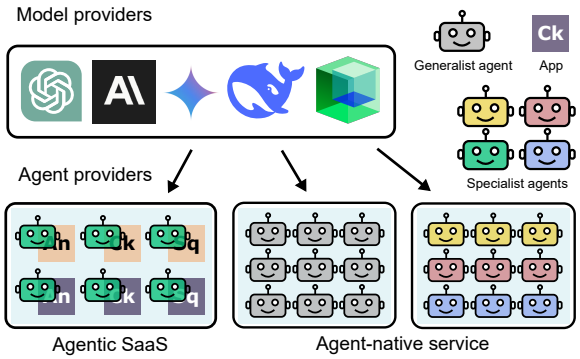

Figure 2: Landscape of the LLM agentic market. LLM agents are provided through agent-native services or agentic SaaS (agents on top of existing software apps). Agent providers are supported by model providers. Colors distinguish agents of different types.

Understanding the current landscape of the agentic market[2] (Fig. 2b and Appendices A-B) is essential for analyzing liability. The present work uses the following terms to refer to its key components: Software platforms offering agent-native services include generalist agents that focuses on general autonomous use of computers, specialist agents which target labor intensive sectors and provide verticalized services for domain-specific workflow automation (Bousetouane, 2025), and character-infused, "hireable" agentic employees. Specialist agents are also offered directly by established software platforms as agentic SaaS to streamline their existing app services. Both agent-native services and agentic SaaS largely source models externally from model providers or derive their own models from open-source projects. Separately, prototypes of integration platforms and integration protocols facilitating the interaction of agents from different frameworks with third-party resources during deployment are also appearing.

Examples of realistic agentic systems exist across many application settings (see Table 1). Because the governance of AI agents remains a nascent topic, potential liability issues in the rapidly expanding agentic market are prevalent but not sufficiently analyzed. Existing efforts are built along two streams: one focuses on establishing and refining the taxonomies of risks and harms using empirical evidence (Chan et al., 2023, 2024; He et al., 2024; Hammond et al., 2025) or differentiating the governance of agentic systems from traditional machine learning (ML) models (Cohen et al., 2024; Kolt, 2025); another focuses on understanding the interactions between humans and AI agents to build constructive principles (Zheng et al., 2023). Research in both streams used principal-agent theory (PAT) (Eisenhardt, 1989; Laffont and Martimort, 2002) as a starting point (see Appendix C), but lacked a systematic examination of how existing legal frameworks for liability can effectively address various principal-agent relationships in AI systems. Moreover, AI systems are typically embedded in a sociotechnical system (Weidinger et al., 2023) such that the verdict on liability issues are finalized only through understanding the agent-environment interactions (see example and analysis in Appendix E). LLM agents are still yet to be mass-deployed, so the liability issues we raise here are based on existing behavioral traits studied in technical research and hypotheticals (see Appendix D.3) extrapolated from them.

Despite the long history of PAT-based legal frameworks (Munday, 2022), their use in AI systems is still in its early days. This work presents an initial attempt to bring principal-agent analysis to current LLM-based agentic systems. Our major contributions include: (i) extending the previous work on PAT for AI governance to accommodate the characteristic behaviors of LLM agents; (ii) discussion of the potential liability issues in light of the emerging market for LLM-based agentic systems; and (iii) outlining of the important directions of policy-driven method developments that support the technical governance of LLM agents. Without loss of generality, our analysis in the present work focuses on US jurisdictions using the Restatement of Torts and the Restatement of Law (see Appendix D) as the starting point to guide interpretation because tort and agency laws are not harmonized across the US. The framework put forward in this article can be applied to other jurisdictions by contextualizing the relevant agent behavior within the corresponding legal frameworks.

2 Related works

AI liability Liability refers to the legal responsibility for one's actions. Legal scholars have treated AI as products and the harm they caused under product liability (Abraham and Rabin, 2019; Buiten, 2024; Sharkey, 2024). In the US, product liability falls under: (i) design defect, (ii) manufacturing defect, and (iii) a warning defect (inadequate instructions) (Abraham, 2012). Anyone harmed by a product can make a claim against any link in the

[2] https://aiagentslist.com/ai-agents-map

supply chain. Usually, product liability is treated under a strict liability rule, courts do not consider the care level that a manufacturer put in place to avoid an accident (see Appendix D for legal definitions). Design defect can be treated under a strict liability rule or an inquiry that also resembles a negligence inquiry, courts consider the care level by considering alternative existing designs.

Service-caused harm usually triggers a negligence inquiry: was "reasonable care" used? If not, the service provider is liable (primary liability) and their principal can also be liable under multiple legal theories. Principals face two primary liabilities: "negligent hiring" for hiring an agent without a reasonable due diligence on its modus operandi; and "negligent supervision" for failing to reasonably monitor or control its agents. Principals face secondary liability, such as vicarious liability (Sykes, 1984; Diamantis, 2023), because of their relationship (when agents acted within the boundaries of employment). A service relationship can also fall under a contractor relationship, which does not trigger vicarious liability. Courts look at the contract and control exercised to categorize the relationship. Below, we assume the principal exercises enough control to trigger vicarious liability.

Software exhibits both "product" (movable goods) and "service" (akin to professional offering) characteristics (Gemignani, 1980; Popp, 2011). As AI agents become more autonomous, they move closer to services, their actions are more accountable due to increasing agenticness (Chopra and White, 2011; Chan et al., 2023). A negligence rule with potential vicarious liability may be more suited (Turner, 2018) for those relationships. In current legal use, this framework straddles product liability falling on the manufacturer and vicarious liability involving principals. Besides the service-product divide, law and economics (L&E) has also advocated for other approaches, including risk-based liability (Geistfeld et al., 2022), fault-based liability (Buiten et al., 2023), explanation-based liability (Padovan et al., 2023), etc.

Principal-agent problems in AI/ML Prior works that invoked the principal-agent framework in AI/ML primarily focused on the decision-making aspects in human-AI collaboration and AI safety. The agents here act as representatives of the principal, which differ from the agents in traditional reinforcement learning (Diaz et al., 2024). Lubars and Tan (2019) discussed the relation between task delegation and the principal's preference. Hadfield-Menell and Hadfield (2019) mapped AI alignment onto the principal-agent problem and discussed the alignment issues in the incomplete contracts theory. Athey et al. (2020) considered different scenarios in allocating decision authority when the human principal and the AI agent have different capabilities. Critch and Krueger (2020) and Hendrycks (2023) considered the potential dangers of complete task delegation to agentic systems. Besides, principal-agent problems have also been considered in game-theoretic machine learning (Gan et al., 2024) and in reinforcement learning settings with two interacting agents (Ivanov et al., 2024).

Agent-oriented software systems Agentic systems have long been proposed as a canonical approach for software design (Jennings, 2001; Zambonelli et al., 2003). This paradigm has experienced further rise in the LLM era (Wang et al., 2024c) because of the role-playing capability (Shanahan et al., 2023; Chen et al., 2024a) of the these models. Their core advantages are scalability, flexibility, and the ability to perform complex tasks through task decomposition. LLM agents (Li, 2025) are configured by text instructions and they can interact with external resources to achieve enhanced capabilities than LLMs in reasoning, tool use, memory, planning, and personalization. LLM-based MASs can use verbal communication protocols to facilitate collaboration and engage in debates (Tran et al., 2025), and the protocol topology is a key for their efficient scaling and behavioral control (Qian et al., 2025).

3 Inherent liability issues in single agents

Contemporary approaches to the governance of AI agents (Kampik et al., 2022; Chan et al., 2023, 2024; Kolt, 2025) resort to PAT, where the principal, the human or company, delegates a task or goal to the AI agent, based on a mutual agreement. Yet LLM agents cannot satisfy all criteria of a normal agent (Perrier and Bennett, 2025) in PAT, creating an agency gap that can lead to an excess of unpredictable actions (John et al., 2025). **Inherent liability issues in agentic systems arise from the dependence structure between the principal and the agent in task delegation as well as the agency gap between LLM agents and normal human agents**. We discuss these issues from the perspective of each key component of PAT (see Section C)

Principal	Agentic system	Delegated tasks
Radiologist	Medical imaging AI agent(s)	Produce a preliminary interpretation of a CT/MRI scan for a patient, suggest additional tests or treatments
Frontend engineer	Website design AI agent(s)	Produce the code for a website from a design specification of each webpage and external media resources
Traveler	Travel booking AI agent(s)	Plan a trip to a series of destinations including the selection of lodgings and transportation vehicles
Shopper	Online shopping AI agent(s)	Seek and aggregate the best online deals that match items from a purchase list within a purchase budget
Insurance policyholder	Insurance claim AI agent(s)	Combine different sources of information (hospital bills, accident reports, personal emails, etc) to draft an insurance claim following a specific format restriction

Table 1: Examples of AI agent use cases cast in the principal-agent framework. Each agentic system can be implemented using a single agent or multiple agents.

in the single-principal and single-agent setting.

3.1 Artificial agency

PAT requires a clarification of the agency relationship, which remains a hotly debated interdisciplinary topic for LLM-based systems (Shavit et al., 2023; Dai, 2024; Barandiaran and Almendros, 2024; Dung, 2024; Perrier and Bennett, 2025; Mattingly and Cibralic, 2025; Butlin, 2025; Das, 2025). AI researchers often take an operational view of artificial agency, such that it is possible to quantify and compare the mental state characteristics of and between AI systems through external interrogation (Baird and Maruping, 2021; Chan et al., 2023; Miehling et al., 2025). Representative caveats of artificial agency for LLMs and LLM agents include:

• **Instability**: Behavior varies with the same or paraphrased but meaning-consistent prompt on different trials (Loya et al., 2023).

• **Inconsistency**: Behavior is sensitive to distracting contextual information or affected by sentiment and adversarially designed prompts (Jain et al., 2023; Maus et al., 2023; Zhuo et al., 2024).

• **Ephemerality**: The complexity of behavioral sequence is restricted by the context window length because of the lack of effective memory mechanisms (Maharana et al., 2024).

• **Planning-limitedness**: Construction of executable plans hinges on accessible environmental feedback which is task-limited (Kambhampati et al., 2024; Wang et al., 2024b; Chen et al., 2024b).

Liability from flawed agency An agency relationship requires both principals' and agents' agreement. Current LLM agents cannot yet form an authentic relationship of such because of their flawed agency (Barandiaran and Almendros, 2024; Perrier and Bennett, 2025). Voluntary relationships usually indicate that both parties benefit. However, because rationality is generally not a built-in goal in developing LLM agents (Macmillan-Scott and Musolesi, 2024), their lack of consent (from agents) means that agent providers may face liability or risks they have not considered. That is why many have consistently argued that AI failures should be treated as product liability and because of its opacity, those failures should be adjudicated under a strict liability rule (Abraham and Rabin, 2019; Buiten, 2024; Sharkey, 2024). AI providers would be left on the hook. This approach would align with the *least-cost avoidance* theory of assigning liability based on who can avoid the same accident at the lowest costs (Calabresi and Melamed, 1972) because users may not be the best place to assess whether the AI agents have any defects that could cause an accident. Even if they can, they may not be able to modify the functioning of the AI agents to avoid an accident.

The problem is that AI providers may not want to be held liable because they do not want to be exposed to unquantifiable risks as they cannot anticipate how the AI users will deploy their AI. In such a situation, the Coase Theorem suggests that the allocation of liability and risks is best left to contracts but before the parties can do that, legislators and courts must clarify rights and liabilities (Coase, 1960). Both the AI providers and users need to use a contract to apportion possible liability or compensation mechanisms.

3.2 Task specification and delegation

A principal delegates tasks to their agents usually either because the principal lacks the resources or the expertise to deal with the tasks (Castelfranchi and Falcone, 1998; Mitchell, 2021). For LLM agents, the principal provides as input to the LLM a task specification that contains instructions on the nature and procedure of task execution, available resources, potential ways to overcome hurdles, the principal's preferences, etc. Such a procedure is subject to the following issues:

• **Task underspecification** is often present because it is impossibly costly to fully anticipate all possible scenarios to put into the specification. Moreover, the same agent may be appropriate for one task but not another; therefore, task specification – the equivalent of job design in organizational theory (Oldham and Fried, 2016) – should balance the tradeoff between agent capability and task complexity (Hadfield-Menell and Hadfield, 2019). This incompleteness leaves the door open to undesirable outcomes such as negative side effects.

• **Risky delegation** The principal tends to forgo the delegation of very high-stakes tasks because of their severe adverse consequences and the unreliability of agent behaviors (Lubars and Tan, 2019). The risk of delegation can be reduced by the amount of repeated feedback the principal provides to ensure alignment (Jiang et al., 2024).

Liability from task misdelegation When selecting the agent (and the tasks) to delegate, principals are expected to carry out their due diligence. If they fail to take reasonable precautions when selecting agents, principals can face liability for negligently selecting (Camacho, 1993). Negligent selection (or negligent hiring) occurs when a principal fails to exercise reasonable care when hiring an agent and the failure caused a third party to suffer harm. Negligent selection applies beyond delegating risky tasks. When the human principal selects the original AI agent, the original selection can trigger a negligent selection. Once the AI agent starts selecting other AI agents to carry out subtasks, that selection is more likely to fall under a theory of negligent supervision (see Section 3.3).

Delegation of tasks also raises information concerns because the principal may not have authorized the agent to share information (Baird and Maruping, 2021). Granting LLM agents access to critical information raises concerns of copyright and trademark (e.g. duplication of protected documents), trade secret (sharing information outside the system), privacy (e.g. transferring data governed by the General Data Protection Regulation or California Consumer Privacy Act), etc. That is why many tasks are not delegable (Mitchell, 2021; Mitchell et al., 2025). A human agent may be able to distinguish between delegable and non-delegable tasks based on the sensitivity of the information, while an AI agent may not without clear instructions (Hadfield-Menell and Hadfield, 2019).

3.3 Principal oversight

Human oversight is a costly endeavor for many AI applications in reality, yet it remains the gold standard in existing AI governance principles (Sterz et al., 2024; Cihon, 2024). Principal-agent problems suffer from information asymmetries, which are usually resolved through monitoring and incentive realignment via reward or punishment. However, monitoring an AI agent requires designs that allow the principal to observe and understand what the agent is doing. Principals who prefer to realign the AI agent's incentives (Everitt et al., 2021) with their own would have to understand what "motivates" these agents or assume that they respond to human-style incentives (Ratliff et al., 2019). At the moment, quantifying misalignment remains challenging, especially for highly capable and general-purpose AI systems (Anwar et al., 2024). Moreover, the quality of principal oversight can be threatened by a spectrum of behaviors LLMs inherit from their training, such as:

• **Sycophancy** refers to the tendency of AI systems to provide responses that the evaluator would prefer in favor of improving the answer (Perez et al., 2023; Sharma et al., 2024), exploiting the evaluator's cognitive biases (e.g. susceptibility to flattery) rather than correctly performing their duty.

• **Manipulation** refers to the ability of LLMs to influence their principals (Campedelli et al., 2024; Burtell and Woodside, 2023; Carroll et al., 2023), towards ends that are non-welfare maximizing to their principals.

• **Deception** refers to the tendency of AI systems to induce false beliefs (Park et al., 2024b; Scheurer et al., 2024; Lang et al., 2024), reinforcing the information asymmetries between the AI agent and its principal.

• **Scheming** refers to the strategic behavior of AI systems to harbor alternative and potentially harm-

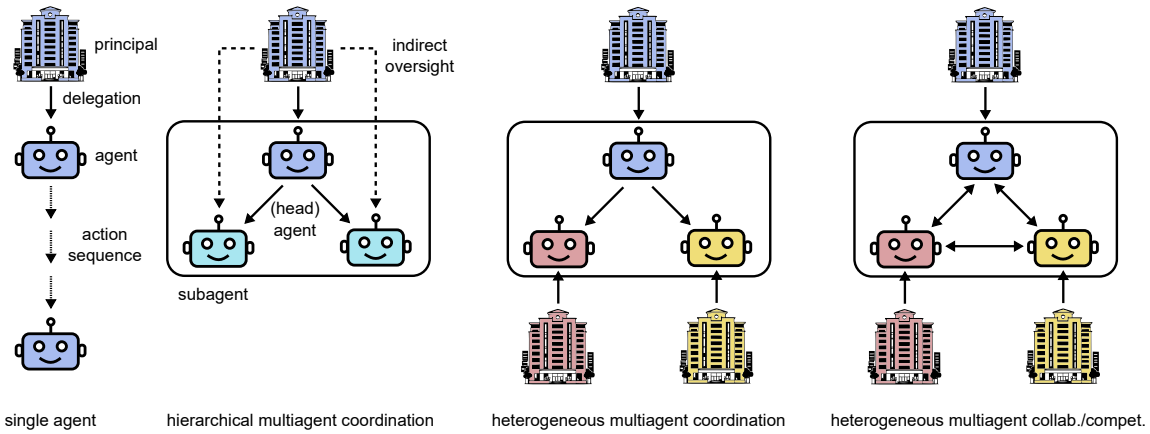

Figure 3: Examples of interaction patterns between the principals and single agents or multiagent systems (MASs). Each MAS in the coordination or collaboration (collab.)/competition (compet.) pattern is enclosed within a box. Distinct principals are colored differently.

ful motives from alignment with the principal during post-training, leading to fake alignment (Greenblatt et al., 2024; Balesni et al., 2024)

These behavioral patterns also manifest in real-world scenarios (Blonz, 2023). All of them can upend the principal-agent asymmetry such that the principal will not be able to reliably monitor the agent's behavior or provide informative feedback.

Liability from compromised oversight Principals that fail to exercise oversight over agents that committed tortious acts are subject to primary liability under a negligent supervision theory (Cavico et al., 2016) act. A principal must take reasonable care to train and supervise their agents. If they fail to do so, they can face liability if the acts (or omissions) of their agents harm third parties or create unreasonable risks. The level of supervision depends on the context. For example, an AI agent that is delegated the tasks of drawing up an architectural plan for a new public library, then making materials selection, and finally ordering those materials needs more supervision than an AI agent that has been put in charge of booking a holiday trip – because the risks of harm differ.

4 Emergent liability issues in MASs

MASs have flexible designs (Fig. 3), so agency relationships and the associated principal-agent problem can occur at different levels, similar to the functioning of a firm (Fama, 1980). The challenges in governing single LLM agents are compounded by the participation of multiple agents in the actions, leading to additional issues with no immediate equivalent in the single-agent setting. **Emergent liability issues arise from coordination in multiagent systems and interactions between agentic systems and with supporting agents in the environment (e.g. an agent platform).** These issues exist in addition to the inherent ones for each agent. An MAS such as in Fig. 1 can contain an orchestrator (agent) that functions as a local manager and directs the work execution of a number of agent workers/teams with different expertise. However, imposing liability only on the orchestrator does not incentivize the improvement of the subagents[3].

4.1 Role and agency allocation

Constructing an LLM-based MAS involves role (or task) allocation (Campbell and Wu, 2011; Guo et al., 2024; Tran et al., 2025), which implicates allocated agency, where the single agents can act on their own to accomplish the goals defined by the assigned task and resources. Because of the role-playing capability (Shanahan et al., 2023; Chen et al., 2024a), current LLM-based MASs generally adopt a role-centric approach, executing role allocation alongside its associated task (Guo et al., 2024; Tran et al., 2025). This approach provides an interpretable division of labor and can directly mimic interactions in teams of humans. Alternatively, role allocation can be self-organized, such as in the deployment of subagents, which also need not be stationary. Role allocation is affected by the nature of LLMs' flawed agency (Section 3.1), leading to potential downstream issues:

• **Influenceability**: Agency of individual agents in an MAS can be enhanced or reduced through communication with other agents in a cooperative

[3]The technical distinction in designing an agentic system with fixed or adaptive roles/tasks for each agent doesn't result in different considerations of liability.

MAS (He et al., 2025), therefore triggering agency shift or unanticipated conduct.

• **Distributedness**: The distribution of agency to different agents in an MAS leads to task specialization and latency which can trade off against performance and speed (Mieczkowski et al., 2025).

• **Diminished control**: In a hierarchical MAS, subagents are more separated from the principal than the head agent, therefore may be harder to directly control or monitor. The principal is more prone to manipulation by the head agent.

Liability from agent misallocation These concerns make task allocation the principal's most important decision. Firstly, the principal must decide which tasks can be delegated and to what type of agents. A principal can face some liability for negligently selecting and supervising. This liability exposure may deter some principal and agentic system deployment – particularly until opacity is not resolved and human-in-the-loop is not optimized. The other associated issue for the principals is the cost-effectiveness. As the number of allocations increases, the agency cost also increases: each new allocation must offer marginal benefits that justify its marginal costs (which also encompasses the liability exposure). From the perspective of L&E, the liability assignment to the entity is favored if the accident can be avoided at the lowest cost (Calabresi and Melamed, 1972; Carbonara et al., 2016).

Secondly, the MAS exposes its principal to the risks their (sub)agents take. The principal must understand the risks associated with each task and with each agent. Then, they must assess which tasks can be delegated without surpassing its risk tolerance. However, the principal faces information asymmetries and might speculate about the expected cost-benefit of each (sub)agent. The principal also faces liability for not being careful in assigning a task to an agent, but usually not for allocating too many tasks (Carbonara et al., 2016). Finally, if the system is so complex or opaque, courts may decide that the harm "speaks for itself", inferring carelessness from the harm — turning a negligence rule into a strict liability rule (Fraser et al., 2022; Casey, 2019). Explainability can therefore become even more necessary.

4.2 Operational uncertainty

As LLM-based MASs become more complex, oversight becomes increasingly challenging. A human overseer may only handle direct communications with the head agent, while the subsequent interactions between the head agent and the subagents are initiated autonomously among themselves. The organizational hierarchy and communication protocol can facilitate the reduction of human involvement, such as in the coordination structures in Fig. 3. Unguided interactions between multiple LLM agents can create complex failure modes depending on the agent architecture and task (Pan et al., 2025), creating additional challenges in their use.

• **Failure cascade** refers to the scenario where the downstream agents can have increased vulnerability than the upstream agents in a MAS, which can be induced by coordination issues (Peigne-Lefebvre et al., 2025) and communication noise induced by a confused agent (Barbi et al., 2025).

• **Agent collusion** refers to the collaboration between agents that negatively impact others (Fish et al., 2024; Lin et al., 2024).

A promising direction for minimizing operational uncertainty from misbehaving agents is to instigate corrective mechanisms and foster a norm-based governance (Hadfield-Menell and Hadfield, 2019; Kampik et al., 2022), where the norm is defined through spontaneous and engineered social interactions between agents (Trivedi et al., 2024).

Liability from operational uncertainty When multiple LLM agents interact autonomously, the attribution of responsibility becomes blurred because decisions emerge from collective behaviors rather than individual actions. A software provider usually bears the liability for the harm caused by their LLM agents:[4] if an AI agent is considered a product that caused harm to a third party, the victim could sue the product manufacturer (i.e., software provider here) under product liability theory, the court would usually apply a strict liability rule and the manufacturer will be held liable if the product is considered to have caused the harm (Turner, 2018; Furr, 2024; Barfield, 2018); if an AI agent is considered a service that caused the harm to a third party, the victim could sue the service provider (i.e., software provider), the court would usually apply a negligence rule (although strict liability may also apply in risky activity e.g., putting an AI in charge of handling dynamites (Reid, 1999)), and

[4]This problem is already present in the autonomous vehicle context and has pushed governments like the UK to change their laws to shield AI providers from liability see UK Automated and Electric Vehicles Act 2018 and UK Automated Vehicles Act 2024 (Soder et al., 2024).

if the service was provided negligently, the victim will recover (Ramakrishnan et al., 2024; Barfield, 2018). The question of whether a piece of software is a product or service remains a question of fact.[5] Attempting to assert that an AI system is a distinct entity has failed in the past and will likely fail in the future (Lior, 2024). This doesn't mean that no one else could be held liable or that the software providers cannot seek contribution under a contract clause, but that they will be usually held liable if their AI agent is deemed to have caused harm to someone to whom they owned a duty of care.

For a heterogeneous MAS, courts may apply separate liabilities or distribute liability among providers according to the harm contribution. Courts do not favor this approach because it is complex to estimate, so the parties would benefit from contractual clarification between all the (sub)agents (e.g. contract liability based on value). Instead, when causes cannot be disentangled and assigned, courts revert to joint and several liability such that each party is liable for the full harm and can be sued individually, which expands liability to the (sub)agents – a tempting approach to avoid complex litigations and battle of experts when MASs are involved (Custers et al., 2025). Courts may decide that multiagent behaviors, such as cascading failures or agent collusion tendencies, lead to third-party harm too often. So, courts may elevate the use of MASs to a "risky activity" and use a strict liability rule (Čerka et al., 2015).

4.3 Platform integration

Emerging LLM-based MASs feature provider-dependent agent frameworks, which will likely follow somewhat different safety protocols (Fig. 3) between agent providers. At the moment, efforts to integrate different agent frameworks analogous to traditional software integration (Bass et al., 2021) are still lagging but are expected to ramp up due to market growth. The motivation to integrate LLM agents is the enhancement of system capability by unifying disparate provider frameworks, which may include privileged access to customized agentic components (e.g. unique databases, fast memory, etc). Overall, integration serves the needs of the user (i.e. principal) by balancing the advantages of different agent frameworks as well as providing an extra layer of control and oversight through the inclusion of what we call *platform agents* (see Fig. 1). At the moment, the potential benefits of integration platforms for LLM agents include:

- **Platform oversight** refers to measures on a platform to provide users with enhanced multiagent security through a security-guard agent (Xiang et al., 2024), collusion mitigation mechanism (Foxabbott et al., 2023), detection and suppression of copyright infringement (Liu et al., 2025a) or privacy leakage. These oversight mechanisms are provided through the integration platform as platform agents that interacts with an existing agentic system.
- **Platform teaming** refers to the formation of agent teams on an integration platform through user-defined or ad hoc protocols (Mirsky et al., 2022; Wang et al., 2024a) that enhance cooperation among homogeneous or heterogeneous agents.

Liability from mismanaged platforms An integrated multiagent platform could carry some liability depending on the level of control it exercises for the digital entities (Gabison and Buiten, 2020; Lefouili and Madio, 2022). For example, control might encompass the behavioral monitoring of the individual agents operating on the platform. Because the platform intrinsically involves multiple principals and multiple agents, the principals may be liable for engaging in collusive behavior but, in rare occasions, platforms have faced liability for incentivizing others (e.g. copyright infringement[6], intentional interference[7]).

5 Policy-driven technical development

The attribution of liability benefits from in-depth failure analysis of system behavior and transparency mechanisms that supports the tracing of agent actions. These in turn motivates concomitant developments in the approaches to manage agent behavior. We discuss a few directions to this end:

Interpretability and behavior evaluations Tracing an AI agent's actions (Lu et al., 2024) can be the basis for establishing whether it took reasonable care (Price et al., 2019; Choi, 2020) and therefore the evidence for liability claims. Prior works on the interpretability and faithfulness in LLM reasoning (Lyu et al., 2023; Wei Jie et al., 2024) and dialogue generation (Tuan et al., 2021) may eventually also be used to assist in the analysis of agent behavior.

[5] See e.g., *Lemmon v. Snap, Inc.*, 995 F. 3d 1085 (9th Cir. 2021); *Holbrook v. Prodomax Automation LTD.*, Case No. 1:17-cv-219 (W.D. Mich. 2021).

[6] *MGM Studios, Inc. v. Grokster, Ltd.*, 545 U.S. 913 (2005)

[7] *hiQ Labs, Inc. v. LinkedIn Corp.*, 938 F. 3d 985 (9th Cir. 2019)

This analysis can be used to investigate whether the AI intended some actions – a necessary element in intentional torts. More generally, these kinds of evaluations may help understand unreliable behavioral patterns and help diagnose and refine the design bottlenecks in LLM agents.

To better ground notions of reasonable care for AI agents, method developments should prioritize decomposing complex multiagent interactions into interpretable causal mechanisms, leveraging causal abstraction frameworks (Geiger et al., 2024) to create faithful, human-intelligible representations of agent interactions that preserve essential causal relationships while abstracting away unnecessary details. Additionally, formal verification approaches (Zhang et al., 2024) may be able to detect and prevent potential failure modes in agent interactions to improve decision-making.

Reward and conflict management Some application settings of LLM-based MASs aim to mimic the functioning of human teams and organizations (Xie et al., 2024). Agentic systems can learn from existing organizational theory (Mitnick, 1992; Vardi and Weitz, 2016) to improve the design and architecture given the flawed agency of its components. The most relevant aspects include systems for managing reward and conflict between agents. It has been shown that "verbal tipping" (Salinas and Morstatter, 2024) can provide concrete incentives in instruction to improve LLM performance. Moreover, managing knowledge conflict in LLMs (Xu et al., 2024) has been the most related area. To manage generic conflicts between LLM agents, a credit system (Thomas et al., 2017) of refusal and sanction based on agent IDs may be beneficial (Chan et al., 2025). This could include adaptive trust scoring from evaluation of domain-specific expertise, and an arbitration protocol to adjudicate conflicts such that the arbiter holds the right to refuse action from a frequently misbehaving agent. While the scoring system would need to be supported across the relevant AI agent integration platforms (Fig. 2a), it is essential to balance individual agent utility and cooperation.

Misalignment and misconduct avoidance As discussed in Section 3.3, LLMs and the agents based on them may act in misaligned ways, with behaviors ranging from sycophancy and deception to scheming. While AI model providers may be able to partially reduce these problems through effective detection and behavior steering, which has been demonstrated on LLMs (Rimsky et al., 2024; Goldowsky-Dill et al., 2025; Williams et al., 2025). In LLM-based agentic systems, the equivalent tasks could be carried out using separate LLM agents through observing and analyzing other agents' behavior, such as using the theory-of-mind capability (Street, 2024). The MAS can include agents with a specially finetuned base model (Binz and Schulz, 2023) as a warden (agent) for deception mitigation. Similar approaches can also be used to suppress other agent misconduct such as the generation of harmful or copyright-protected content using the warden to filter through key tokens and phrases that can induce such behavior. Recently, Hua et al. (2024) implemented a safety inspection step with a specialized LLM agent to improve operational safety. Such adaptive approach carried out by a separate agent can compensate for the limitations in existing model-level approaches such as machine unlearning (Bourtoule et al., 2021), which is suffering from limited effectiveness and tradeoffs with other model capabilities (Cooper et al., 2024; Liu et al., 2025b).

6 Conclusion

We examined liability issues arising from LLM-based agentic systems by analyzing and situating distinct aspects of the agentic AI ecosystem according to the principal-agent theory. Although a varieties of issues of LLM agents are yet to present themselves in concrete real-world examples, the growing evidence and demonstrations in simulated scenarios can inform their potential impact at the societal scale, which we built on in our prospective study. Our work shows that besides increasing agency, disruption in other aspects of the principal-agent relationship can also lead to liability incidents. Ultimately, the materialization of liability issues in reality will be dominated by the incidents that occur in the more frequent use cases of agentic systems in each industry sector. Our analysis enriches existing contextualization of AI risk (Chan et al., 2023; Hammond et al., 2025) and demonstrates the explanatory power of the behavior-centric approach to translating frontier AI research into tangible knowledge to inform legal analysis and policy.

Limitations

The present work focused on legal liabilities but did not address the potential moral responsibility

issues present in LLM-based agentic systems. Although liability depends on the legal system and therefore varies across jurisdictions, the principle of vicarious liability for the actions of an agent remains consistent in most jurisdictions. The discussions were set for relatively small-scale agentic systems (e.g. up to tens of agents), where each agent has its detailed roles and tasks. Because of the execution cost and reliability issues, we don't think larger-scale MASs (e.g. having a hundred or more agents) will become practical solutions and deployed widely or in long-running instances any time soon. Their use tends to be more relevant for academic research and they have different operating conditions, including more common role/task underspecification and stronger emergent characteristics. Answering questions about failure modes and attributing liability would require more understanding of the deployment history. Another limitation is that our discussion centered around LLM agents, but the same liability issues described here are likely applicable to multimodal AI agents.

Ethics statements

The design and deployment of LLM-based agentic systems raise significant ethical and legal considerations regarding liability attribution. This work acknowledges the complex interplay between provider responsibility, user actions, and the emergent behaviors of semi-autonomous systems powered by LLMs. We recognize that traditional liability frameworks may inadequately address scenarios where AI agents make consequential decisions with incomplete human oversight. Our analysis aims to contribute to the discussion on appropriate liability for all sides related to LLM-based agentic systems that balances current evidence from system behavior and the available legal framework. We have considered liability in different deployment environments with the aim of informing policy discussions and technical developments to make liability more traceable in AI systems that are complex and extensible, but prone to misbehavior and failure and lack in self-control.

Acknowledgments

We thank the organizers of the AI Governance at the Crossroads symposium held at Berkeley, California in February, 2025. We thank Micah Carroll at the University of California, Berkeley for information about LLM behaviors and AI safety and helpful comments on the manuscript.

References

Kenneth S Abraham. 2012. *The Forms and Functions of Tort Law*. Foundation Press.

Kenneth S Abraham and Robert L Rabin. 2019. Automated vehicles and manufacturer responsibility for accidents. *Virginia Law Review*, 105(1):127–171.

George A Akerlof. 1970. The market for "lemons": Quality uncertainty and the market mechanism. *The Quarterly Journal of Economics*, 84:488–500.

American Law Institute. 1965. *Restatement (Second) of Torts*. American Law Institute Publishers, Philadelphia.

American Law Institute. 2006. *Restatement (Third) of Agency*. American Law Institute Publishers, Philadelphia.

American Law Institute. 2010. *Restatement (Third) of Torts*. American Law Institute Publishers, Philadelphia. Various volumes, e.g., Liability for Physical and Emotional Harm, Products Liability.

Usman Anwar, Abulhair Saparov, Javier Rando, Daniel Paleka, Miles Turpin, Peter Hase, Ekdeep Singh Lubana, Erik Jenner, Stephen Casper, Oliver Sourbut, Benjamin L. Edelman, Zhaowei Zhang, Mario Günther, Anton Korinek, Jose Hernandez-Orallo, Lewis Hammond, Eric J. Bigelow, Alexander Pan, Lauro Langosco, and 23 others. 2024. Foundational Challenges in Assuring Alignment and Safety of Large Language Models. *Transactions on Machine Learning Research*.

Kenneth J Arrow. 1963. Uncertainty and the welfare economics of medical care. *The American Economic Review*, 53:941–973.

Susan C. Athey, Kevin A. Bryan, and Joshua S. Gans. 2020. The Allocation of Decision Authority to Human and Artificial Intelligence. *AEA Papers and Proceedings*, 110:80–84.

Aaron Baird and Likoebe Maruping. 2021. The Next Generation of Research on IS Use: A Theoretical Framework of Delegation to and from Agentic IS Artifacts. *Management Information Systems Quarterly*, 45(1):315–341.

Mikita Balesni, Marius Hobbhahn, David Lindner, Alexander Meinke, Tomek Korbak, Joshua Clymer, Buck Shlegeris, Jérémy Scheurer, Charlotte Stix, Rusheb Shah, Nicholas Goldowsky-Dill, Dan Braun, Bilal Chughtai, Owain Evans, Daniel Kokotajlo, and Lucius Bushnaq. 2024. Towards evaluations-based safety cases for AI scheming. *arXiv preprint*. ArXiv:2411.03336 [cs].

Xabier E. Barandiaran and Lola S. Almendros. 2024. Transforming Agency. On the mode of existence of Large Language Models. *arXiv preprint*. ArXiv:2407.10735 [cs].

Ohav Barbi, Ori Yoran, and Mor Geva. 2025. Preventing Rogue Agents Improves Multi-Agent Collaboration. *arXiv preprint*. ArXiv:2502.05986 [cs].

Woodrow Barfield. 2018. Liability for autonomous and artificially intelligent robots. *Paladyn, Journal of Behavioral Robotics*, 9(1):193–203.

Len Bass, Paul Clements, and Rick Kazman. 2021. *Software Architecture in Practice*. Addison-Wesley Professional.

Umang Bhatt, Sanyam Kapoor, Mihir Upadhyay, Ilia Sucholutsky, Francesco Quinzan, Katherine M. Collins, Adrian Weller, Andrew Gordon Wilson, and Muhammad Bilal Zafar. 2025. When Should We Orchestrate Multiple Agents? *arXiv preprint*. ArXiv:2503.13577 [cs].

Marcel Binz and Eric Schulz. 2023. Turning large language models into cognitive models. In *The Twelfth International Conference on Learning Representations*.

Joshua A. Blonz. 2023. The Costs of Misaligned Incentives: Energy Inefficiency and the Principal-Agent Problem. *American Economic Journal: Economic Policy*, 15(3):286–321.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. 2021. Machine unlearning. In *2021 IEEE symposium on security and privacy (SP)*, pages 141–159. IEEE.

Fouad Bousetouane. 2025. Agentic Systems: A Guide to Transforming Industries with Vertical AI Agents. *arXiv preprint*. ArXiv:2501.00881 [cs].

Miriam Buiten, Alexandre de Streel, and Martin Peitz. 2023. The law and economics of AI liability. *Computer Law & Security Review*, 48:105794.

Miriam C. Buiten. 2024. Product liability for defective AI. *European Journal of Law and Economics*, 57(1):239–273.

Matthew Burtell and Thomas Woodside. 2023. Artificial influence: An analysis of ai-driven persuasion. *arXiv preprint arXiv:2303.08721*.

Patrick Butlin. 2025. The agency in language agents. *Inquiry*, 0(0):1–21. Publisher: Routledge _eprint: https://doi.org/10.1080/0020174X.2024.2439995.

Guido Calabresi and A Douglas Melamed. 1972. Property rules, liability rules, and inalienability: one view of the cathedral. *Havard Law Review*, 85:1089–1128.

Rodolfo A Camacho. 1993. How to avoid negligent hiring litigation. *Whittier Law Review*, 14:787–808.

Adam Campbell and Annie S. Wu. 2011. Multi-agent role allocation: issues, approaches, and multiple perspectives. *Autonomous Agents and Multi-Agent Systems*, 22(2):317–355.

Gian Maria Campedelli, Nicolò Penzo, Massimo Stefan, Roberto Dessì, Marco Guerini, Bruno Lepri, and Jacopo Staiano. 2024. I want to break free! persuasion and anti-social behavior of llms in multi-agent settings with social hierarchy. *arXiv preprint arXiv:2410.07109*.

Emanuela Carbonara, Alice Guerra, and Francesco Parisi. 2016. Sharing residual liability: the cheapest cost avoider revisited. *The Journal of Legal Studies*, 45(1):173–201.

Micah Carroll, Alan Chan, Henry Ashton, and David Krueger. 2023. Characterizing Manipulation from AI Systems. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–13, New York, NY, USA. ACM.

Bryan Casey. 2019. Robot Ipsa Loquitur. *Georgetown Law Journal*, 108(2):225–286.

Cristiano Castelfranchi and Rino Falcone. 1998. Towards a theory of delegation for agent-based systems. *Robotics and Autonomous Systems*, 24(3):141–157.

Frank J Cavico, Bahaudin G Mujtaba, Marissa Samuel, and Stephen C Muffler. 2016. The tort of negligence in employment hiring, supervision, and retention. *American Journal of Business and Society*, 1(4):205.

Alan Chan, Carson Ezell, Max Kaufmann, Kevin Wei, Lewis Hammond, Herbie Bradley, Emma Bluemke, Nitarshan Rajkumar, David Krueger, Noam Kolt, Lennart Heim, and Markus Anderljung. 2024. Visibility into AI Agents. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pages 958–973, New York, NY, USA. Association for Computing Machinery.

Alan Chan, Rebecca Salganik, Alva Markelius, Chris Pang, Nitarshan Rajkumar, Dmitrii Krasheninnikov, Lauro Langosco, Zhonghao He, Yawen Duan, Micah Carroll, Michelle Lin, Alex Mayhew, Katherine Collins, Maryam Molamohammadi, John Burden, Wanru Zhao, Shalaleh Rismani, Konstantinos Voudouris, Umang Bhatt, and 3 others. 2023. Harms from Increasingly Agentic Algorithmic Systems. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, pages 651–666, New York, NY, USA. Association for Computing Machinery.

Alan Chan, Kevin Wei, Sihao Huang, Nitarshan Rajkumar, Elija Perrier, Seth Lazar, Gillian K. Hadfield, and Markus Anderljung. 2025. Infrastructure for AI Agents. *arXiv preprint*. ArXiv:2501.10114 [cs].

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, Aili Chen, Nianqi Li, Lida Chen, Caiyu Hu, Siye Wu, Scott Ren, Ziquan Fu, and Yanghua

Xiao. 2024a. From Persona to Personalization: A Survey on Role-Playing Language Agents. *Transactions on Machine Learning Research*.

Yanan Chen, Ali Pesaranghader, Tanmana Sadhu, and Dong Hoon Yi. 2024b. Can We Rely on LLM Agents to Draft Long-Horizon Plans? Let's Take TravelPlanner as an Example. *arXiv preprint*. ArXiv:2408.06318 [cs].

Bryan H. Choi. 2020. Software as a Profession. *Harvard Journal of Law & Technology (Harvard JOLT)*, 33(2):557–638.

Samir Chopra and Laurence F. White. 2011. *A Legal Theory for Autonomous Artificial Agents*. University of Michigan Press, Ann Arbor.

Peter Cihon. 2024. Chilling autonomy: Policy enforcement for human oversight of AI agents. In *41st International Conference on Machine Learning, Workshop on Generative AI and Law*.

Ronald Harry Coase. 1960. The problem of social cost. *The Journal of Law and Economics*, 3:1–44.

Michael K. Cohen, Noam Kolt, Yoshua Bengio, Gillian K. Hadfield, and Stuart Russell. 2024. Regulating advanced artificial agents. *Science*, 384(6691):36–38. Publisher: American Association for the Advancement of Science.

A. Feder Cooper, Christopher A. Choquette-Choo, Miranda Bogen, Matthew Jagielski, Katja Filippova, Ken Ziyu Liu, Alexandra Chouldechova, Jamie Hayes, Yangsibo Huang, Niloofar Mireshghallah, Ilia Shumailov, Eleni Triantafillou, Peter Kairouz, Nicole Mitchell, Percy Liang, Daniel E. Ho, Yejin Choi, Sanmi Koyejo, Fernando Delgado, and 16 others. 2024. Machine Unlearning Doesn't Do What You Think: Lessons for Generative AI Policy, Research, and Practice. *arXiv preprint*. ArXiv:2412.06966 [cs].

N. Criado, E. Argente, and V. Botti. 2011. Open issues for normative multi-agent systems. *AI Communications*, 24(3):233–264. Publisher: IOS Press.

Andrew Critch and David Krueger. 2020. AI Research Considerations for Human Existential Safety (ARCHES). *arXiv preprint*. ArXiv:2006.04948 [cs].

Bart Custers, Henning Lahmann, and Benjamyn I Scott. 2025. From liability gaps to liability overlaps: shared responsibilities and fiduciary duties in ai and other complex technologies. *AI & SOCIETY*, pages 1–16.

Jessica Dai. 2024. Position: Beyond Personhood: Agency, Accountability, and the Limits of Anthropomorphic Ethical Analysis. In *Forty-first International Conference on Machine Learning*.

Parashar Das. 2025. Agency in Artificial Intelligence Systems. *arXiv preprint*. ArXiv:2502.10434 [cs].

Mihailis E. Diamantis. 2023. Vicarious Liability for AI. *Indiana Law Journal*, 99(1):317–334.

Manfred Diaz, Joel Z. Leibo, and Liam Paull. 2024. Milnor-Myerson Games and The Principles of Artificial Principal-Agent Problems. In *Finding the Frame: An RLC Workshop for Examining Conceptual Frameworks*.

Leonard Dung. 2024. Understanding Artificial Agency. *The Philosophical Quarterly*, page pqae010.

Kathleen M. Eisenhardt. 1989. Agency Theory: An Assessment and Review. *The Academy of Management Review*, 14(1):57–74. Publisher: Academy of Management.

Tom Everitt, Ryan Carey, Eric D. Langlois, Pedro A. Ortega, and Shane Legg. 2021. Agent Incentives: A Causal Perspective. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(13):11487–11495. Number: 13.

Eugene F. Fama. 1980. Agency Problems and the Theory of the Firm. *Journal of Political Economy*, 88(2):288–307. Publisher: The University of Chicago Press.

Eugene F. Fama and Michael C. Jensen. 1983. Agency Problems and Residual Claims. *The Journal of Law & Economics*, 26(2):327–349. Publisher: [University of Chicago Press, Booth School of Business, University of Chicago, University of Chicago Law School].

Sara Fish, Yannai A. Gonczarowski, and Ran I. Shorrer. 2024. Algorithmic Collusion by Large Language Models. *arXiv preprint*. ArXiv:2404.00806 [econ].

Jack Foxabbott, Sam Deverett, Kaspar Senft, Samuel Dower, and Lewis Hammond. 2023. Defining and Mitigating Collusion in Multi-Agent Systems.

Henry Fraser, Rhyle Simcock, and Aaron J Snoswell. 2022. Ai opacity and explainability in tort litigation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 185–196.

Jessica Furr. 2024. The untouchables: Why software companies escape liability for faulty software? *The LawVerse Substack*.

Garry A. Gabison and Miriam C. Buiten. 2020. Platform Liability in Copyright Enforcement. *Columbia Science and Technology Law Review*, 21(2):237–281.

Jiarui Gan, Minbiao Han, Jibang Wu, and Haifeng Xu. 2024. Generalized Principal-Agency: Contracts, Information, Games and Beyond. *arXiv preprint*. ArXiv:2209.01146 [cs].

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. 2024. Causal Abstraction: A Theoretical Foundation for Mechanistic Interpretability. *arXiv preprint*. ArXiv:2301.04709 [cs].

Mark A. Geistfeld, Ernst Karner, and Bernhard A. Koch. 2022. Comparative Law Study on Civil Liability for Artificial Intelligence. In Ernst Karner, Bernhard A. Koch, Mark A. Geistfeld, and Christiane Wendehorst, editors, *Civil Liability for Artificial Intelligence and Software*, pages 1–184. De Gruyter.

Michael C. Gemignani. 1980. Product Liability and Software. *Rutgers Computer & Technology Law Journal*, 8(2):173–204.

Nicholas Goldowsky-Dill, Bilal Chughtai, Stefan Heimersheim, and Marius Hobbhahn. 2025. Detecting Strategic Deception Using Linear Probes. *arXiv preprint*. ArXiv:2502.03407 [cs].

Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Sam Marks, Johannes Treutlein, Tim Belonax, Jack Chen, David Duvenaud, Akbir Khan, Julian Michael, Sören Mindermann, Ethan Perez, Linda Petrini, Jonathan Uesato, Jared Kaplan, Buck Shlegeris, Samuel R. Bowman, and Evan Hubinger. 2024. Alignment faking in large language models. *arXiv preprint*. ArXiv:2412.14093 [cs].

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large Language Model Based Multi-agents: A Survey of Progress and Challenges. volume 9, pages 8048–8057. ISSN: 1045-0823.

Dylan Hadfield-Menell, Mckane Andrus, and Gillian Hadfield. 2019. Legible Normativity for AI Alignment: The Value of Silly Rules. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, pages 115–121, New York, NY, USA. Association for Computing Machinery.

Dylan Hadfield-Menell and Gillian K. Hadfield. 2019. Incomplete contracting and ai alignment. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, page 417–422, New York, NY, USA. Association for Computing Machinery.

Lewis Hammond, Alan Chan, Jesse Clifton, Jason Hoelscher-Obermaier, Akbir Khan, Euan McLean, Chandler Smith, Wolfram Barfuss, Jakob Foerster, Tomáš Gavenčiak, The Anh Han, Edward Hughes, Vojtěch Kovařík, Jan Kulveit, Joel Z. Leibo, Caspar Oesterheld, Christian Schroeder de Witt, Nisarg Shah, Michael Wellman, and 25 others. 2025. Multi-Agent Risks from Advanced AI. *arXiv preprint*. ArXiv:2502.14143 [cs].

Pengfei He, Yupin Lin, Shen Dong, Han Xu, Yue Xing, and Hui Liu. 2025. Red-Teaming LLM Multi-Agent Systems via Communication Attacks. *arXiv preprint*. ArXiv:2502.14847 [cs].

Yifeng He, Ethan Wang, Yuyang Rong, Zifei Cheng, and Hao Chen. 2024. Security of AI Agents. *arXiv preprint*. ArXiv:2406.08689 [cs].

Dan Hendrycks. 2023. Natural Selection Favors AIs over Humans. *arXiv preprint*. ArXiv:2303.16200 [cs].

Bengt Holmström. 1979. Moral hazard and observability. *The Bell journal of economics*, 10:74–91.

Wenyue Hua, Xianjun Yang, Mingyu Jin, Zelong Li, Wei Cheng, Ruixiang Tang, and Yongfeng Zhang. 2024. TrustAgent: Towards Safe and Trustworthy LLM-based Agents. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10000–10016, Miami, Florida, USA. Association for Computational Linguistics.

Dima Ivanov, Paul Dütting, Inbal Talgam-Cohen, Tonghan Wang, and David C. Parkes. 2024. Principal-Agent Reinforcement Learning: Orchestrating AI Agents with Contracts. *arXiv preprint*. ArXiv:2407.18074 [cs].

Shrey Jain, Zoë Hitzig, and Pamela Mishkin. 2023. Contextual Confidence and Generative AI. *arXiv preprint*. ArXiv:2311.01193 [cs].

Nicholas R. Jennings. 2001. An agent-based approach for building complex software systems. *Commun. ACM*, 44(4):35–41.

Michael C. Jensen and William H. Meckling. 1976. Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of Financial Economics*, 3(4):305–360.

Song Jiang, Da Ju, Andrew Cohen, Sasha Mitts, Aaron Foss, Justine T. Kao, Xian Li, and Yuandong Tian. 2024. Towards Full Delegation: Designing Ideal Agentic Behaviors for Travel Planning. In *NeurIPS 2024 Workshop on Adaptive Foundation Models*.

Sotiropoulos John, Rosario Ron F. Del, Kokuykin Evgeniy, Oakley Helen, Habler Idan, Underkoffler Kayla, Huang Ken, Steffensen Peter, Aralimatti Rakshith, Bitton Ron, Sharbat Tamir Ishay, Giarrusso Vinnie, Kutal Volkan, Howe Allie, Bhartiya Anshuman, Sheriff Akram, Guilherme Emmanuel, Rogers Eric, Martin Itsik, and 67 others. 2025. OWASP Top 10 for LLM Apps & Gen AI Agentic Security Initiative. Technical report, OWASP.

Subbarao Kambhampati, Karthik Valmeekam, Lin Guan, Mudit Verma, Kaya Stechly, Siddhant Bhambri, Lucas Paul Saldyt, and Anil B. Murthy. 2024. Position: LLMs Can't Plan, But Can Help Planning in LLM-Modulo Frameworks. In *Forty-first International Conference on Machine Learning*.

Timotheus Kampik, Adnane Mansour, Olivier Boissier, Sabrina Kirrane, Julian Padget, Terry R. Payne, Munindar P. Singh, Valentina Tamma, and Antoine Zimmermann. 2022. Governance of Autonomous Agents on the Web: Challenges and Opportunities. *ACM Trans. Internet Technol.*, 22(4):104:1–104:31.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae Won Park. 2024. MDAgents: An Adaptive Collaboration of LLMs for Medical Decision-Making. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Noam Kolt. 2025. Governing AI Agents. *arXiv preprint*. ArXiv:2501.07913 [cs].

Iuliia Kotseruba and John K. Tsotsos. 2020. 40 years of cognitive architectures: core cognitive abilities and practical applications. *Artificial Intelligence Review*, 53(1):17–94.

Naveen Krishnan. 2025. AI Agents: Evolution, Architecture, and Real-World Applications. *arXiv preprint*. ArXiv:2503.12687 [cs].

Jean-Jacques Laffont and David Martimort. 2002. *The Theory of Incentives: The Principal-Agent Model*. Princeton University Press, Princeton, N.J.

Leon Lang, Davis Foote, Stuart Russell, Anca Dragan, Erik Jenner, and Scott Emmons. 2024. When Your AIs Deceive You: Challenges of Partial Observability in Reinforcement Learning from Human Feedback. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Yassine Lefouili and Leonardo Madio. 2022. The economics of platform liability. *European Journal of Law and Economics*, 53(3):319–351.

Saul Levmore. 1990. Probabilistic recoveries, restitution, and recurring wrongs. *The Journal of Legal Studies*, 19(S2):691–726.

Xinzhe Li. 2025. A Review of Prominent Paradigms for LLM-Based Agents: Tool Use, Planning (Including RAG), and Feedback Learning. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9760–9779, Abu Dhabi, UAE. Association for Computational Linguistics.

Julien Libert. 2025. An AI Liability Regulation would complete the EU's AI strategy. *CEPS*.

Ryan Y. Lin, Siddhartha Ojha, Kevin Cai, and Maxwell Chen. 2024. Strategic Collusion of LLM Agents: Market Division in Multi-Commodity Competitions. In *Language Gamification - NeurIPS 2024 Workshop*.

Anat Lior. 2024. Holding ai accountable: Addressing the ai-related harms through existing tort doctrines. *U. Chi. L. Rev. Online*, page 1.

Shunchang Liu, Zhuan Shi, Lingjuan Lyu, Yaochu Jin, and Boi Faltings. 2025a. CopyJudge: Automated Copyright Infringement Identification and Mitigation in Text-to-Image Diffusion Models. *arXiv preprint*. ArXiv:2502.15278 [cs].

Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Yuguang Yao, Chris Yuhao Liu, Xiaojun Xu, Hang Li, Kush R. Varshney, Mohit Bansal, Sanmi Koyejo, and Yang Liu. 2025b. Rethinking machine unlearning for large language models. *Nature Machine Intelligence*, 7(2):181–194. Publisher: Nature Publishing Group.

Manikanta Loya, Divya Sinha, and Richard Futrell. 2023. Exploring the Sensitivity of LLMs' Decision-Making Capabilities: Insights from Prompt Variations and Hyperparameters. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3711–3716, Singapore. Association for Computational Linguistics.

Jiaying Lu, Bo Pan, Jieyi Chen, Yingchaojie Feng, Jingyuan Hu, Yuchen Peng, and Wei Chen. 2024. AgentLens: Visual Analysis for Agent Behaviors in LLM-based Autonomous Systems. *IEEE Transactions on Visualization and Computer Graphics*, pages 1–17. Conference Name: IEEE Transactions on Visualization and Computer Graphics.

Brian Lubars and Chenhao Tan. 2019. Ask not what AI can do, but what AI should do: Towards a framework of task delegability. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. Faithful Chain-of-Thought Reasoning. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 305–329, Nusa Dua, Bali. Association for Computational Linguistics.

Olivia Macmillan-Scott and Mirco Musolesi. 2024. (Ir)rationality and cognitive biases in large language models. *Royal Society Open Science*, 11(6):240255. Publisher: Royal Society.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 2024. Evaluating Very Long-Term Conversational Memory of LLM Agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13851–13870, Bangkok, Thailand. Association for Computational Linguistics.

James Mattingly and Beba Cibralic. 2025. *Machine Agency*. MIT Press.

Natalie Maus, Patrick Chao, Eric Wong, and Jacob R. Gardner. 2023. Black Box Adversarial Prompting for Foundation Models. In *The Second Workshop on New Frontiers in Adversarial Machine Learning*.

Elizabeth Mieczkowski, Ruaridh Mon-Williams, Neil Bramley, Christopher G. Lucas, Natalia Velez, and Thomas L. Griffiths. 2025. Predicting Multi-Agent Specialization via Task Parallelizability. *arXiv preprint*. ArXiv:2503.15703 [cs].

Erik Miehling, Karthikeyan Natesan Ramamurthy, Kush R. Varshney, Matthew Riemer, Djallel Bouneffouf, John T. Richards, Amit Dhurandhar, Elizabeth M. Daly, Michael Hind, Prasanna Sattigeri, Dennis Wei, Ambrish Rawat, Jasmina Gajcin, and

Werner Geyer. 2025. Agentic AI Needs a Systems Theory. *arXiv preprint*. ArXiv:2503.00237 [cs].

Reuth Mirsky, Ignacio Carlucho, Arrasy Rahman, Elliot Fosong, William Macke, Mohan Sridharan, Peter Stone, and Stefano V. Albrecht. 2022. A Survey of Ad Hoc Teamwork Research. In *Multi-Agent Systems*, pages 275–293, Cham. Springer International Publishing.

Margaret Mitchell, Avijit Ghosh, Alexandra Sasha Luccioni, and Giada Pistilli. 2025. Fully Autonomous AI Agents Should Not be Developed. *arXiv preprint*. ArXiv:2502.02649 [cs].

Neil J. Mitchell. 2021. *Why Delegate?* Oxford University Press, New York, NY.

Barry M. Mitnick. 1992. The Theory of Agency and Organizational Analysis. In *Ethics and Agency Theory: An Introduction*. Oxford University Press, Oxford, United Kingdom.

Roderick Munday. 2022. *Agency: Law and Principles*, 4th edition. Oxford University Press, Oxford, United Kingdom ; New York, NY.

Greg R. Oldham and Yitzhak Fried. 2016. Job design research and theory: Past, present and future. *Organizational Behavior and Human Decision Processes*, 136:20–35.

Paulo Henrique Padovan, Clarice Marinho Martins, and Chris Reed. 2023. Black is the new orange: how to determine AI liability. *Artificial Intelligence and Law*, 31(1):133–167.

Melissa Z. Pan, Mert Cemri, Lakshya A. Agrawal, Shuyi Yang, Bhavya Chopra, Rishabh Tiwari, Kurt Keutzer, Aditya Parameswaran, Kannan Ramchandran, Dan Klein, Joseph E. Gonzalez, Matei Zaharia, and Ion Stoica. 2025. Why Do Multiagent Systems Fail? In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*.

Joon Sung Park, Carolyn Q. Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S. Bernstein. 2024a. Generative Agent Simulations of 1,000 People. *arXiv preprint*. ArXiv:2411.10109 [cs].

Peter S. Park, Simon Goldstein, Aidan O'Gara, Michael Chen, and Dan Hendrycks. 2024b. AI deception: A survey of examples, risks, and potential solutions. *Patterns*, 5(5). Publisher: Elsevier.

Pierre Peigne-Lefebvre, Mikolaj Kniejski, Filip Sondej, Matthieu David, Jason Hoelscher-Obermaier, Christian Schroeder de Witt, and Esben Kran. 2025. Multi-Agent Security Tax: Trading Off Security and Collaboration Capabilities in Multi-Agent Systems. *arXiv preprint*. ArXiv:2502.19145 [cs].

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Benjamin Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, and 44 others. 2023. Discovering Language Model Behaviors with Model-Written Evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13387–13434, Stroudsburg, PA, USA. Association for Computational Linguistics.

Elija Perrier and Michael Timothy Bennett. 2025. Position: Stop Acting Like Language Model Agents Are Normal Agents. *arXiv preprint*. ArXiv:2502.10420 [cs].

Jinghua Piao, Yuwei Yan, Jun Zhang, Nian Li, Junbo Yan, Xiaochong Lan, Zhihong Lu, Zhiheng Zheng, Jing Yi Wang, Di Zhou, Chen Gao, Fengli Xu, Fang Zhang, Ke Rong, Jun Su, and Yong Li. 2025. AgentSociety: Large-Scale Simulation of LLM-Driven Generative Agents Advances Understanding of Human Behaviors and Society. *arXiv preprint*. ArXiv:2502.08691 [cs].

Karl Popp. 2011. Software Industry Business Models. *IEEE Software*, 28(4):26–30. Conference Name: IEEE Software.

W. Nicholson Price, Sara Gerke, and I. Glenn Cohen. 2019. Potential Liability for Physicians Using Artificial Intelligence. *JAMA*, 322(18):1765.

Chen Qian, Zihao Xie, YiFei Wang, Wei Liu, Kunlun Zhu, Hanchen Xia, Yufan Dang, Zhuoyun Du, Weize Chen, Cheng Yang, Zhiyuan Liu, and Maosong Sun. 2025. Scaling Large Language Model-based Multi-Agent Collaboration. In *The Thirteenth International Conference on Learning Representations*.

Ketan Ramakrishnan, Gregory Smith, and Conor Downey. 2024. U.S. Tort Liability for Large-Scale Artificial Intelligence Damages: A Primer for Developers and Policymakers. Technical report, RAND Corporation.

Lillian J. Ratliff, Roy Dong, Shreyas Sekar, and Tanner Fiez. 2019. A Perspective on Incentive Design: Challenges and Opportunities. *Annual Review of Control, Robotics, and Autonomous Systems*, 2(Volume 2, 2019):305–338. Publisher: Annual Reviews.

Elspeth Reid. 1999. Liability for Dangerous Activities: A Comparative Analysis. *International & Comparative Law Quarterly*, 48(4):731–756.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. Steering Llama 2 via Contrastive Activation Addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, Bangkok, Thailand. Association for Computational Linguistics.

Abel Salinas and Fred Morstatter. 2024. The Butterfly Effect of Altering Prompts: How Small Changes

and Jailbreaks Affect Large Language Model Performance. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4629–4651, Bangkok, Thailand. Association for Computational Linguistics.

Jérémy Scheurer, Mikita Balesni, and Marius Hobbhahn. 2024. Large Language Models can Strategically Deceive their Users when Put Under Pressure. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role play with large language models. *Nature*, 623(7987):493–498.

Catherine M. Sharkey. 2024. A Products Liability Framework for AI. *Columbia Science and Technology Law Review*, 25(2):240–260. Number: 2.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna M. Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2024. Towards Understanding Sycophancy in Language Models. In *The Twelfth International Conference on Learning Representations*.

Yonadav Shavit, Sandhini Agarwal, Miles Brundage, Steven Adler Cullen O'Keefe, Rosie Campbell, Teddy Lee, Pamela Mishkin, Tyna Eloundou, Alan Hickey, Katarina Slama, Lama Ahmad, Paul McMillan, Alex Beutel, Alexandre Passos, and David G. Robinson. 2023. Practices for Governing Agentic AI Systems.

Lisa Soder, Julia Smakman, Connor Dunlop, Weiwei Pan, Siddharth Swaroop, and Noam Kolt. 2024. Levels of Autonomy: Liability in the age of AI Agents. In *Workshop on Socially Responsible Language Modelling Research*.

Michael Spence. 1973. Job market signaling. *The Quarterly Journal of Economics*, 87:355–374.

Sarah Sterz, Kevin Baum, Sebastian Biewer, Holger Hermanns, Anne Lauber-Rönsberg, Philip Meinel, and Markus Langer. 2024. On the Quest for Effectiveness in Human Oversight: Interdisciplinary Perspectives. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pages 2495–2507, New York, NY, USA. Association for Computing Machinery.

Winnie Street. 2024. LLM Theory of Mind and Alignment: Opportunities and Risks. *arXiv preprint*. ArXiv:2405.08154 [cs].

Theodore Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas Griffiths. 2023. Cognitive Architectures for Language Agents. *Transactions on Machine Learning Research*.

Alan Sykes. 1984. The Economics of Vicarious Liability. *Yale Law Journal*. Accepted: 2021-11-26T12:26:37Z.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2024. MedAgents: Large Language Models as Collaborators for Zero-shot Medical Reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 599–621, Bangkok, Thailand. Association for Computational Linguistics.

Lyn Thomas, Jonathan Crook, and David Edelman. 2017. *Credit Scoring and Its Applications*, second edition. Mathematics in Industry. Society for Industrial and Applied Mathematics.

Khanh-Tung Tran, Dung Dao, Minh-Duong Nguyen, Quoc-Viet Pham, Barry O'Sullivan, and Hoang D. Nguyen. 2025. Multi-Agent Collaboration Mechanisms: A Survey of LLMs. *arXiv preprint*. ArXiv:2501.06322 [cs].

Rakshit Trivedi, Nikhil Chandak, Andrei Ioan Muresanu, Shuhui Zhu, Atrisha Sarkar, Joel Z. Leibo, Dylan Hadfield-Menell, and Gillian K. Hadfield. 2024. Altared Environments: The Role of Normative Infrastructure in AI Alignment. In *Agentic Markets Workshop at ICML 2024*.

Yi-Lin Tuan, Connor Pryor, Wenhu Chen, Lise Getoor, and William Yang Wang. 2021. Local Explanation of Dialogue Response Generation. In *Advances in Neural Information Processing Systems*, volume 34, pages 404–416. Curran Associates, Inc.

Jacob Turner. 2018. *Robot Rules: Regulating Artificial Intelligence*. Palgrave Macmillan, New York, NY.

Yoav Vardi and Ely Weitz. 2016. *Misbehavior in Organizations: A Dynamic Approach*, 2nd edition. Routledge.

Caroline Wang, Arrasy Rahman, Ishan Durugkar, Elad Liebman, and Peter Stone. 2024a. N-agent Ad Hoc Teamwork. *Advances in Neural Information Processing Systems*, 37:111832–111862.

Qian Wang, Tianyu Wang, Zhenheng Tang, Qinbin Li, Nuo Chen, Jingsheng Liang, and Bingsheng He. 2025. All It Takes Is One Prompt: An Autonomous LLM-MA System. In *ICLR 2025 Workshop on Foundation Models in the Wild*.

Xingyao Wang, Yangyi Chen, Lifan Yuan, Yizhe Zhang, Yunzhu Li, Hao Peng, and Heng Ji. 2024b. Executable code actions elicit better LLM agents. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *ICML'24*, pages 50208–50232, Vienna, Austria. JMLR.org.

Yanlin Wang, Wanjun Zhong, Yanxian Huang, Ensheng Shi, Min Yang, Jiachi Chen, Hui Li, Yuchi Ma, Qianxiang Wang, and Zibin Zheng. 2024c. Agents in Software Engineering: Survey, Landscape, and Vision. *arXiv preprint*. ArXiv:2409.09030 [cs].

Yeo Wei Jie, Ranjan Satapathy, Rick Goh, and Erik Cambria. 2024. How Interpretable are Reasoning Explanations from Prompting Large Language Models? In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2148–2164, Mexico City, Mexico. Association for Computational Linguistics.

Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, Iason Gabriel, Verena Rieser, and William Isaac. 2023. Sociotechnical Safety Evaluation of Generative AI Systems. *arXiv preprint*. ArXiv:2310.11986 [cs].

Marcus Williams, Micah Carroll, Adhyyan Narang, Constantin Weisser, Brendan Murphy, and Anca Dragan. 2025. On Targeted Manipulation and Deception when Optimizing LLMs for User Feedback. In *The Thirteenth International Conference on Learning Representations*.

Zhen Xiang, Linzhi Zheng, Yanjie Li, Junyuan Hong, Qinbin Li, Han Xie, Jiawei Zhang, Zidi Xiong, Chulin Xie, Carl Yang, Dawn Song, and Bo Li. 2024. GuardAgent: Safeguard LLM Agents by a Guard Agent via Knowledge-Enabled Reasoning. *arXiv preprint*. ArXiv:2406.09187 [cs].

Yijia Xiao, Edward Sun, Di Luo, and Wei Wang. 2025. TradingAgents: Multi-Agents LLM Financial Trading Framework. In *The First MARW: Multi-Agent AI in the Real World Workshop at AAAI 2025*.

Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Shiyang Lai, Kai Shu, Jindong Gu, Adel Bibi, Ziniu Hu, David Jurgens, James Evans, Philip Torr, Bernard Ghanem, and Guohao Li. 2024. Can Large Language Model Agents Simulate Human Trust Behavior? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. Knowledge Conflicts for LLMs: A Survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8541–8565, Miami, Florida, USA. Association for Computational Linguistics.

Franco Zambonelli, Nicholas R. Jennings, and Michael Wooldridge. 2003. Developing multiagent systems: The Gaia methodology. *ACM Trans. Softw. Eng. Methodol.*, 12(3):317–370.

Yedi Zhang, Yufan Cai, Xinyue Zuo, Xiaokun Luan, Kailong Wang, Zhe Hou, Yifan Zhang, Zhiyuan Wei, Meng Sun, Jun Sun, Jing Sun, and Jin Song Dong. 2024. The Fusion of Large Language Models and Formal Methods for Trustworthy AI Agents: A Roadmap. *arXiv preprint*. ArXiv:2412.06512 [cs].

Qingxiao Zheng, Zhongwei Xu, Abhinav Choudhry, Yuting Chen, Yongming Li, and Yun Huang. 2023. Synergizing Human-AI Agency: A Guide of 23 Heuristics for Service Co-Creation with LLM-Based Agents. *arXiv preprint*. ArXiv:2310.15065 [cs].

Jingming Zhuo, Songyang Zhang, Xinyu Fang, Haodong Duan, Dahua Lin, and Kai Chen. 2024. ProSA: Assessing and Understanding the Prompt Sensitivity of LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1950–1976, Miami, Florida, USA. Association for Computational Linguistics.

Paulius Čerka, Jurgita Grigienė, and Gintarė Sirbikytė. 2015. Liability for damages caused by artificial intelligence. *Computer Law & Security Review*, 31(3):376–389.

A Agentic system definitions

We clarify here the definitions and scopes for some of the key terms used throughout the text. They are not meant for a complete characterization of these terms but are primarily aimed at illustrating their relationships in the context of this work.

• An **AI agent** is a generic term referring to a software agent powered by any form of artificial intelligence (Krishnan, 2025). An equivalent definition is provided in Section 1.

• An **LLM agent** refers specifically to an AI agent powered by at least one LLM as the central component that executes planning, initiates and coordinates the agent's actions, etc. An **LLM-based multiagent system** constitutes of multiple LLM agents that can be derived from the same or different LLMs.

• An LLM-based MAS consists of some components that facilitate its operation. This includes **agent teams** that consist of several interacting agents, which mimics human communication in teamwork and group decision-making. An **orchestrator** is an agent that distributes tasks to other specialized agents and facilitates their actions (Bhatt et al., 2025). On an agent or integration platform, the platform maintainers can order specialized agents, or **platform agents**, to oversee and inspect (Hua et al., 2024) the communications between agents. They are generally aimed at improving the safety (Xiang et al., 2024), security, and policy compliance of the actions carried out on the platform.

• An **AI system** is an umbrella term defined in the EU AI Act[8]. When deployed, an AI system can operate with various levels of autonomy and can

[8] https://artificialintelligenceact.eu/article/3/

create recommendations and content, make predictions and decisions that influence the environment. The system in the term indicates that the AI is not acting in isolation, but is assisted by the surrounding infrastructure, such as cloud computing, databases, and user interfaces, which are integral to the AI system and essential for its use. An AI system can be agentic or non-agentic.

• An **agentic system**, also known as an **agenic AI system** (Shavit et al., 2023), is a type of AI system that contains a level of agency to carry out actions on its own in pursuit of a goal. An agentic system can include a single agent or multiple agents acting in coordination, competition, cooperation (or collaboration). An AI agent is a key component of an agentic system.

• Analogous to the previous entry, **agentic market** is the segment of AI market represented by the providers and buyers of agentic AI systems.

• An **agent platform** provides resources and toolkits to construct, configure as well as deploy AI agents (Fig. 2a). AI agents from distinct providers can operate on a (software) **integration platform**, where they can interact with each other, with third-party data sources, proprietary APIs, etc. AI agents on any deployment platform may be subject to compliance governance and can receive protection against cyberthreats or malfunction from the software infrastructure there.

B Types and examples of existing LLM agent providers

We present here preliminary examples for elements of the agentic market that are currently available.

B.1 Agentic software as a service (SaaS)

• Salesforce: www.salesforce.com/agentforce/

• Adobe: business.adobe.com/products/experience-platform/agent-orchestrator.html

• SAP: www.sap.com/products/artificial-intelligence/ai-agents.html

• Oracle: www.oracle.com/artificial-intelligence/generative-ai/agents/

• Cisco Webex: www.webex.ai/ai-agent.html

B.2 Agent-native service

Providers of generalist agents

• OpenAI Operator: operator.chatgpt.com

• Google DeepMind Project Astra: deepmind.google/technologies/project-astra

• Manus: manus.im

• Simular: simular.ai

Providers of specialist agents

• Sesame (www.sesame.com) offers voice AI agents for different domains.

• Contextual (contextual.ai) offers specialized AI agents with advanced retrieval-augmented features.

• Devin (devin.ai) offers AI agents for coding.

• Sierra (sierra.ai) offers AI agents tailored for different types of customer services.

• Health Force (www.healthforce.ai) offers human resources AI agents to handle digital text processing tasks in healthcare systems.

• Zenity (www.zenity.io/) offers security-focused AI agents.

Providers of character-infused agents

• Artisan: www.artisan.co

• Sintra: sintra.ai

B.3 Elements of agent integration platforms

• Microsoft Copilot Studio: www.microsoft.com/microsoft-copilot/microsoft-copilot-studio

• Anthropic Model Context Protocol: modelcontextprotocol.io

• Google Agent2Agent protocol: google-a2a.github.io/A2A/

• IBM Bee AI: beeai.dev

C Principal agency and liability

PAT examines the relationship where one party, the principal, delegates authority to another, the agent, creating three fundamental challenges (Eisenhardt, 1989; Laffont and Martimort, 2002):

• **Adverse selection** (aka. hidden information problem) occurs when agents possess more information than principals about their abilities or efforts. Adverse selection is a type of information asymmetry also known as the hidden information problem (Akerlof, 1970).

• **Moral hazard** (aka. hidden action problem) occurs when agents take greater risks than principals would prefer because they do not bear the full consequences (Arrow, 1963). Moral hazard is also a type of information asymmetry.

• **Misaligned interest** (aka. conflict of interest) can manifest between principals and agents in four

ways (Jensen and Meckling, 1976): principal-agent collusion against third parties, principal-third party collusion against agents, agent-third party collusion against principals, or agents simply pursuing self-interest independently. Misaligned interests lead to agency cost, which is associated with information sharing, monitoring of the agent, etc (Fama and Jensen, 1983).

Each of these problems has legal and economic solutions. Legal solutions include fiduciary duties (duty of care, duty of loyalty, etc.). Those duties provide principals with a legal recourse against agents who breached those duties. However, legal enforcement is probabilistic and slow (Levmore, 1990), so principals often prefer to use economic mechanisms, which depend on the problem. Principals can address some hidden information problems by creating separating equilibria that force agents to reveal information about themselves, that is, "signals"; signals include credentials, warranties, or performance histories that distinguish high-quality from low-quality agents (Spence, 1973). Principals mitigate a hidden action problem through monitoring (i.e. direct observation of agent behavior) and then linking compensation to observable effort, or bonding arrangements in which the principals tie remuneration to outcomes (bonuses), thereby aligning financial incentives (Holmström, 1979). A principal can address conflicts of interest by realigning incentives through carefully designed contracts (deferred compensation), creating organizational structures that promote incentive realignment (profit sharing), or leveraging reputation mechanisms. A principal has incentives to monitor their agents.

The primary liability faced by agents can motivate them to take reasonable care when carrying out a task. The question is whether AI agents or their providers face any liability in such a system. Primary and secondary liability exposures incentivize principals to ensure that their agents complete the tasks with reasonable care. Overall, the ultimate principal is always a human and may be held liable for the actions of the AI agents.

D Legal definitions and hypotheticals

D.1 Tort law-related terms

In simple terms, **tort law** is the set of laws that attempt to redress civil wrongdoings which ensue from the actions (or omissions) of an individual that cause harm to a third party. Tort law depends on the jurisdiction (i.e., jurisdiction-specific). In the United States, tort law remains the domain of the states. In the European Union, each member state sets its own tort law. A recent effort to harmonize tort law has failed to progress in the European Parliament because of the inability of member states to come to an agreement (Libert, 2025). The definitions below are based on the various Restatements of Tort Law. These Restatements are treatises written by American scholars and published by the American Law Institute. They summarize tort concepts and guide courts and lawyers in the area. The various Restatements (American Law Institute, 1965, 2010) are not binding laws but provide a good indication of what courts consider in the United States.

• Tort law does not require the actor and the third party to have any relationship or a legally recognized privity (which, by contrast, is required for contract claims). Every time an individual (or entity) acts, they create the chance of harming someone. The **intent** of the actor is generally not relevant in tort law to decide whether someone is responsible for most torts. However, intent can affect whether to impose punitive damages, which are damages that go beyond compensating for the harm caused.[9]

• In tort law, **liability** refers to a court finding that an individual or entity is responsible for the harm inflicted on a third party. To prove that an individual (or entity) is liable, the victim must provide different evidence depending on the rule. The most common rules are strict liability, the negligence rule, and comparative negligence.

• The **strict liability rule** is a type of liability rule where the victim-cum-plaintiff must demonstrate to a judge (or jury) that the action-taker-cum-defendant took actions and that the actions caused the defendant harm that can be redressed usually either through a monetary compensation called damages or a court order called injunction to stop their harmful actions. The strict liability rule ignores the level of care the defendant took in avoiding the accident. Courts usually use a strict liability for abnormally dangerous activities[10] or for many product liability questions [11]

[9] *Restatement (Third) of Torts: Liability for Physical and Emotional Harm § 1*

[10] *Restatement (Third) of Torts: Liability for Physical and Emotional Harm Ch. 4*

[11] *Restatement (Third) of Torts: Product Liability § 1 1998.*

• The **negligence rule** is a type of liability rule where the victim-cum-plaintiff must demonstrate to a judge (or jury) that the action-taker-cum-defendant had a duty toward the victim, failed to live up to that duty (i.e., did not take reasonable care), and that their actions caused the victim redressable harm. The negligence rule attempts to ensure that potential tortfeasors take reasonable efforts to lower the risk imposed on others. For example, "an actor ordinarily has a duty to exercise reasonable care when the actor's conduct creates a risk of physical harm."[12]

• When **comparative negligence rules** are used, courts weigh the responsibility of both the plaintiff and the defendant in causing the accident and assign to each litigant a percentage of responsibility for the accident. Most courts reduce the recoverable damages by the percentage of fault attributed to the plaintiff (an approach known as "pure comparative negligence") or prevent all recovery if the plaintiff is more at fault than the defendant ("modified or partial comparative negligence").

• **Product liability** refers to the liability traced from the harm caused by a product back to its manufacturer. Product liability occurs when a product is defective and "A product is defective when, at the time of sale or distribution, it contains a manufacturing defect, is defective in design or is defective because of inadequate instructions or warnings."[13] The manufacturing defect occurred because the normal manufacturing process was not followed and this deviation led to the product presenting harm, while a product manufactured using the normal process does not. The deviation affects only one product (or a batch of products) that has been put into the stream of commerce.

By contrast, a design defect and a warning defect affect all the products that have been put into the stream of commerce. A design defect occurred when the manufacturer created a product with a feature that presents more risk than benefit. A common test to assess whether a design defect occurred is the risk-utility test, which requires the plaintiff to show that the manufacturer could have used a reasonable alternative design, that the alternative design is safer than the marketed product, and that the alternative design was reasonably economically feasible and practical. Finally, a warning defect or information defect occurs because certain hidden risks cannot be designed away, so the manufacturer must provide information to warn the consumers, but the information was not adequate because it was not visible, comprehensive, only used an icon, and no risk mitigation information was provided.

D.2 Agency law-related terms

In simple terms, agency law governs the duties and obligations that arise in an agency relationship. Once again, the definitions below rely on the various Restatements of Law, including the Restatement of Agency (American Law Institute, 2006) because agency law is specific to the jurisdiction.

• An **agency relationship** is established when the "principal" (an individual or entity) agrees that the "agent" acts (another individual or entity) on their behalf, the principal can control the agent's activities, and the agent agrees to the relationship.[14]

• In many cases, **control** is often the key factor that courts investigate to identify whether two individuals entered into an agency relationship. Courts do not need to find that the principal exercised control and only need to find that the principal had the ability to do so. Even if the principal exercises control and provides detailed instructions, the agent may not act in the way the principal expected because agents must interpret those instructions.

The existence of an agency relationship triggers various types of liabilities. The ones discussed above include:

• **Negligent hiring** is a form of primary liability — because it is based on the actions of the principal — where the principal is held responsible for selecting an agent that was likely to cause harm to a third party and put them in a position that creates the chance to cause that harm.[15] For example, a store manager could be held liable for negligent hiring for hiring someone known for excessive drinking as a delivery person who ends up assaulting customers in their home.[16]. The principal has a duty to exercise reasonable care when selecting an employee to act on their behalf.

• **Negligent supervision** is a form of primary liability where the principal is held responsible for failing to provide training or supervise an agent who is found to have caused harm to a third party.[17]

[12] *Restatement (Third) of Torts: Liability for Physical and Emotional Harm § 7(a)*

[13] *Restatement (Third) of Torts: Product Liability § 2*

[14] *Restatement (Third) of Agency (2006) § 1.01*

[15] *Restatement (Third) of Agency § 7.03 (2006)*

[16] *Fleming v. Bronfin*, 80 A.2d 915 (D.C. 1951)

[17] *Restatement (Third) of Agency § 7.03 (2006)*

For example, an auction house could be held liable for hiring ex-convicts to serve as security guards and and failing to control them when they forcefully removed customers from the premises and harming them in the process.[18]

• **Vicarious liability** is a form of secondary liability – because the principal is responsible for the actions of another person based on the relationship between the principal and the agent – where the agent caused harm to a third party within the scope of the agency relationship[19]. The scope of the relationship defines the boundaries of the principal's responsibility. If the principal provides some instructions to the agent and the agents carries them to the letter and then causes harm to a third party, the principal is held vicariously liable and the third party can sue the agent and the principal. However, even if the agent deviates from the principal's instructions, the principal may still be held vicariously liable. For example, Company A hires 3 taxi drivers to work in 8-hour shifts in New York City. Driver X hit Pedestrian Y and broke Pedestrian Y's leg while carrying a fare from Wall Street to Times Square. Pedestrian sues Driver X and Company A to recover his medical bills and the pain and suffering from having a leg problem.[20]

D.3 Hypotheticals

The following examples of liability analysis are written from a legal perspective using hypothetical names. They contrast with those discussed in the main text sections 4-5 using examples largely from machine learning, language models and agents.

Agentic home security system Acme corp. offers private security system services to independent homeowners. When hired, the company sends employees to install the security system into a home including components such as cameras, infrared, electric fences, etc. Aside from installation, Acme relies on an AI agent with the ability to trigger various systems (e.g., automated sprinklers) to monitor homes and intrusions. In case of an intrusion, Acme's AI agent can contact the police or private security for a welfare check or deploy a drone to conduct a scan of the property. The drone surveillance is provided by comp. developing AI-powered drones. The comp. uses its own system to control the drones and carry out security checks. All those communications are AI to AI and no human comes into the decision loop whether to trigger a sprinkler or send a drone, but some humans may participate in the act during checks (e.g., police officers).

In 2025, Peter Principal hired Acme corp. to provide security to its chemical factory. A few months later, Ted Thieve breaks into the factory. Acme's AI agent detects the intrusion with its infrared sensors installed on the periphery of the factory and asks AI drone to send a drone to provide aerial pictures of the factory to confirm the intrusion. Before it could even reach Peter Principal's factory. The drone loses control and drops on Victoria Victim's car, damaging it. Victoria Victim sues Peter Principal, Acme corp. and AI-powered drone to recover the damages caused to her car.

Agentic package delivery system ABC corp. offers services for delivery systems. They claim to be able to deliver any package weighing two pounds or under to any of the 48 contiguous states within 24 hours. To be able to do so, ABC corp. uses a sophisticated agentic system that decides the route, hires local delivery services and handles the packages. In other words, its system coordinates with and contract various service providers. Many of the local service providers use their own agentic system to decide what deliver service to accept, which to reject, the route, and the costing. The agentic systems do more than logistics. In many cases, the agentic system also uses delivery robots to deliver the packages and those delivery robots are controlled by those same agentic system. All those communications are AI to AI and no human comes into the decision loop, but some humans may take part in delivering service (e.g., truck driver).

In 2025, Peter Principal hired ABC corp. to deliver a new birthday cake from New York City, New York to Salt Lake City, Utah. ABC's agentic system accepted the package. It contacted Delivery Express's agentic system to fly the package from New York City to Utah and Drone Delivery to take the package from the airport to the delivery address. During the flight, the AI-powered drone delivery service has a malfunction and drops the package on Victoria Victim, who is injured. Victoria Victim sues Peter Principal, ABC corp., Delivery Express, and Drone Delivery to recover the medical bills and the pain and suffering.

[18] *American Auto. Auction, Inc. v. Titsworth*, 730 S.W.2d 499 (1987)

[19] *Restatement (Third) of Agency § 7.05 (2006)*

[20] *Restatement (Third) of Agency 7.07 Employee Acting Within Scope of Employment*

E Principal-agent analysis of an LLM-related legal case

The case features a few different principal-agent relations, starting with Roberto Mata, the plaintiff. The segment relevant for the current discussion concerns the law firm that hired the two lawyers representing Mata. A part of the tasks that the lawyers do was delegated to the LLM, ChatGPT, which is the subagent of the law firm. The law firm can control the tools that the lawyers use through the employment contract, but ChatGPT was *not* explicitly excluded. The lawyers defended the fake cases generated by ChatGPT and were therefore ruled by the judge to be in bad faith. The law firm also received punishment alongside its employees. The case of Mata v. Avianca, Inc. (above) was much discussed in the public media in 2023 as an early incident involving the use of LLMs. The delegational structure of major entities involved in the case is (Fig. 4)

$$\begin{aligned}\text{Roberto Mata} &\rightarrow \text{Law firm}\\ &\rightarrow \text{Mata's lawyers}\\ &\rightarrow \text{ChatGPT}\end{aligned}$$

Mata v. Avianca, Inc.

In 2023, a US Federal Judge reprimanded two lawyers and their law firm for acting in bad faith and making misleading statements to the court. Their crime? They trusted ChatGPT.[21] *ChatGPT made up cases, the lawyers failed to notice, but the judge did. The judge ordered the lawyers to produce the cases. The lawyers "doubled down and did not begin to dribble out the truth" for another few weeks. The court punished the lawyers for the actions of ChatGPT and their law firm in the process. ChatGPT (and its provider OpenAI) did not bear any responsibility because the lawyers were the ones who had a duty to provide accurate information to the court. The lawyers responded that: "We made a good faith mistake in failing to believe that a piece of technology could be making up cases out of whole cloth." These lawyers and their principal learned not to trust AIs.*

The prompting method that the lawyers used was not explicitly discussed in the court proceeding. ChatGPT is treated as a subagent that carries out the lawyers' task request. It was largely regarded as an instance of LLM hallucination (or confabulation) in the news media rather than deception. The judicial verdict was finalized based on the behavior of the lawyers. The legal case highlights the shortcomings of current legal frameworks and the technical gap to fully comprehend the behaviors of AI systems in the presence of their environment

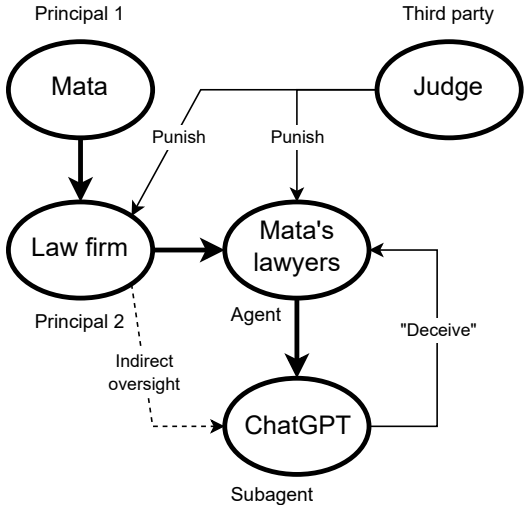

Figure 4: Principal-agent analysis of Mata vs. Avianca, Inc.

(involving humans) and make use of this evidence in attributing blame.

[21] *Mata v. Avianca, Inc.*, 678 F. Supp. 3d 443 (SDNY 2023)