AGENTS AREN'T AGENTS: THE AGENCY, LOYALTY AND ACCOUNTABILITY PROBLEMS OF AI AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

The rapid adoption of AI agents marks a shift from predictable digital services to systems entrusted with autonomous, judgment-like tasks. As people delegate more responsibility to these agents, questions of control, loyalty, and accountability become urgent. Yet today's agents are operated through fragmented layers of control by developers, hosts, and providers, which blur lines of responsibility and divide loyalties before users ever interact with them. Without reconsideration, we risk misallocating responsibility, overstating loyalty, and obscuring who ultimately benefits from these systems. In this paper, we systematically discuss key issues that hinder AI agents from attaining true legal agency. We identify three unresolved problems: **Agency**—who is the principal and who is the agent in the polyadic governance of AI development and deployment; **Loyalty**—whether AI agents can serve the principal's best interests; and **Accountability**—when AI agents make mistakes, who is responsible for them? We examine the technological foundations that give rise to these problems and highlight key limitations of the current agency law framework in addressing emerging issues related to AI agents. As a position paper, our study offers fresh perspectives on AI agents from a legal standpoint and could inspire new research directions in this domain.

1 INTRODUCTION

AI agents are rapidly transitioning from demos to daily use.[1] Consumers rely on them to draft messages, plan travel, track deliveries, and organize meetings. Firms deploy them to field support inquiries, summarize contracts, triage operations tickets, and trigger back-office actions through APIs. Developers embed agentic components that monitor mailboxes, watch data streams, and initiate workflows without being manually invoked. Systems that act *for* a user are becoming part of routine consumer and enterprise processes (Bengio et al., 2025; South et al., 2025; Kolt, 2025). These systems are not just another interface on top of traditional software. An agent accepts a *goal*, chooses the means, and adapts to new information (e.g., rebooking when a flight is delayed or following up when a supplier does not respond). They interact in natural language with people and services, compose multi-step plans, call tools, and persist over time.

This functional shift naturally leads people to reach for familiar legal analogies, especially the law of agency, to resolve misalignments between AI agents and human users (Lior, 2019; Benthall & Shekman, 2023; Koessler, 2024; Riedl & Desai, 2025; Kolt, 2025). In the human setting, an agency relationship arises when a Principal manifests that an Agent shall act on the Principal's behalf, the Agent consents, and the Principal retains a right of control. Doctrine then allocates authority (actual and apparent), imposes fiduciary duties (loyalty, obedience, care, and candor), and assigns liabilities among Principals, Agents, and Third Parties.

However, invoking this framework carries risks. For consumers, the analogy invites misplaced reliance, leading them to assume an AI agent is their legal Agent, bound by loyalty and responsibility.

[1] In this paper, we use *AI agents* as a technical term of convenience to denote AI systems, typically a large language model integrated with tools, that can pursue goals, decompose tasks, and act for or on behalf of a user. When referring to the legal categories defined in agency law, we capitalize Agent, Principal, and Third Party to avoid conflating functional delegation with legal status.

For scholars and policymakers, it suggests that agency law can straightforwardly resolve alignment problems, when in fact the structural differences between human and AI agents make a transplant misleading. Table 1 outlines four of the most common misconceptions about AI agents and agency law, explaining why they appear plausible and why they collapse under closer examination.

Misconception	Why it seems plausible	Reality
An AI agent is my Agent.	Conversational interfaces and tool use mimic human assistance; commercial branding invites the mental model of a personal Agent.	AI systems lack legal personhood and responsibility. The legal Agent is the party that deploys the system. Interactions are triadic, with providers shaping outcomes that the user cannot control.
An AI agent is more loyal than humans.	Models have no self-interest, are persistent/always-on, and do not 'get tempted,' so they appear more faithful to user goals.	Loyalty is structurally divided. Multiple rule-imposers (trainers, hosts, providers, and users) bind behavior. *Undivided* loyalty to a single Principal is impossible unless safeguards are overridden (the 'AI henchman' risk).
Applying Agency law makes AI agents loyal.	Liability disciplines human Agents; by analogy, legal pressure should yield faithful performance.	Liability cannot discipline models directly. Model behaviors are significantly different from humans and can have disloyal behaviors in unexpected ways.
AI agents owe fiduciary duties like human agents.	State courts require all types of agents to bear duties to users.	Agency duties are modifiable by contract and vary by state. Providers often use ToS (arbitration, class waivers, liability caps) to narrow remedies. Unlike licensed professions, most AI services lack external discipline or non-waivable obligations.

Table 1: Common misconceptions about AI agents and Agency law

These discrepancies stem from the *anthropocentric* nature of agency law. As Cohen (2019) notes, fiduciary duties presuppose personal relationships, mutual intelligibility, and "human rhythms" of interaction. Agency doctrine disciplines self-governing, self-interested human agents by constraining their natural tendency to pursue their own advantage at the expense of their principal. It deters betrayal through fiduciary duties and liability, while at the same time protecting third parties who rely on the agent's representation. AI agents, on the other hand, have no self-preserving motives or reputational stakes. AI agents have, if any, only *engineered autonomy* follwing the rules imposed by multiple actors, from trainers to providers to users (Feng et al., 2025). This *polyadic* nature of governance prevents AI agents to provide *undivided loyalty* to a single user, the atmost premise in the agency law. As AI agents cannot feel deterrence in response to liability or reputational loss, the incentive structures that discipline human agents cannot directly correct AI behavior.

In this position paper, we argue that treating AI systems as if they were human Agents obscures fundamental structural differences in how they are built, operated, and governed. We frame three core challenges—**Agency**, **Loyalty**, and **Accountability**—that emerge from the polyadic nature of AI governance and the inability of current systems to provide undivided loyalty or bear responsibility. Rather than attempting a doctrinal transplant from human agency law, we diagnose the technological and institutional foundations that produce these gaps and highlight the limitations of existing legal frameworks in addressing them. Our goal is not to resolve these questions conclusively, but to surface them as central to the future governance of AI agents and to offer a foundation for legal scholarship, policy design, and interdisciplinary research.

2 How AI Agents differ from existing digital services

In this section, we discuss why AI agents are essentially different from existing digital services. Unlike traditional digital services that operate in a relatively fixed environment, AI agents are designed to be more autonomous and can act on behalf of the user to perform various tasks in real-world environments. They are goal-driven systems that can plan, select tools, and adapt to changing information streams, rather than simply executing a pre-coded sequence of steps. Because these systems

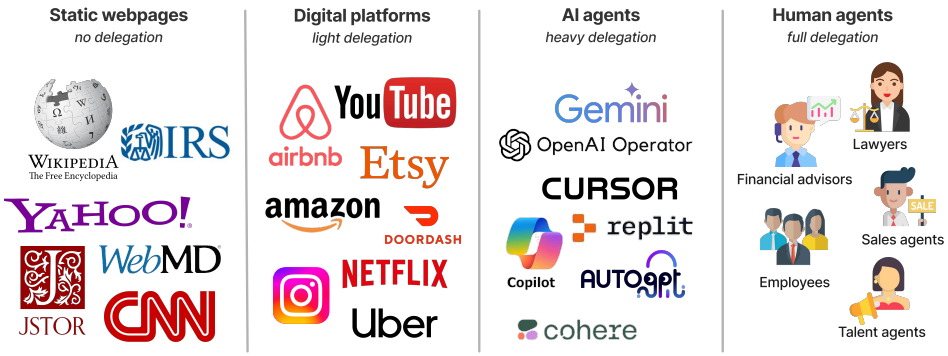

Figure 1: Users have a higher level of delegation for AI agents, making them more similar to human agents instead of existing digital services.

largely occupy roles that look like *acting for* a person. Making choices, communicating with others, and taking consequential steps, they invite comparison with human agents in the legal sense.

Delegation. With AI agents, users delegate outcomes, not just clicks Zhu et al. (2025); Guggenberger et al. (2023). Instead of instructing a service to "open page A, then press button B," a user expresses a *goal*—e.g, "reschedule my afternoon meetings around a 3 pm dentist appointment", and the agent decomposes the given objectives into tasks, queries calendars, drafts messages to attendees, and proposes new times. This is qualitatively different from conventional software, which requires the user to specify every intermediate action. AI agents also operate under constraints ("don't cancel with client X," "stay under $200," or "use my company account"), requiring complex reasoning. The practical effect is that the locus of decision-making shifts from the user's hands to the agent's planning layer, making the delegation relationship both more efficient and less transparent.

Interactivity. AI agents do not act in isolation; they interact with a variety of parties and systems in fluid, conversational ways Muller & Weisz (2022); Wan et al. (2024); Borghoff et al. (2025); OpenAI (2025). They are able to send emails or chats in natural language, negotiate meeting times, call APIs, and exchange structured data with platforms for payments, bookings, and support tickets. They can maintain context over time, remember preferences, and adjust tone or strategy based on feedback; for example, softening a collection's message after a recipient responds defensively, or escalating a customer support issue when a scripted workflow stalls. This interactive capacity means agents can create expectations and induce reliance in third parties (e.g., issuing confirmations, placing holds, or making representations), which starts to resemble how human agents create practical commitments on behalf of principals.

Autonomy. AI agents are expected to act with higher autonomy, aligning three key axes: initiative, adaptation, and persistence Liu et al. (2023); Feng et al. (2025); Hughes et al. (2025). *Initiative* appears when agents trigger themselves based on events ("if a high-priority email arrives, draft a response and propose a call"). *Adaptation* emerges when they revise plans in light of new information (a flight delay prompts rebooking and hotel changes without being told step-by-step what to do). *Persistence* shows up in long-running workflows that span days or weeks, where the agent monitors states, retries, and follows up. It is *engineered autonomy* to choose means toward user-specified ends under uncertainty. Still, the overall behavior is functionally agentic: selecting actions, balancing constraints, and affecting the user's legal and practical position.

Together, the properties of delegation, interactivity, and autonomy distinguish AI agents from traditional digital services that are usually operated within a certain scope Lanham (2025). They act *for* someone, *with* others, and *on* the world. Therefore, AI agents should be discussed through the lens of human agency instead of existing algorithms or digital services. Appendix A.3 provides a more detailed overview of the changes in digital services, AI agents, and human Agents across delegation, interactivity, and autonomy.

3 WHAT IS LEGAL AGENCY?

Category	Key Elements
Fiduciary Duties	*Undivided loyalty:* Act solely for the principal, not for self or conflicting third parties. No multiple principals unless all consent. (§§ 3.14–3.15, §§ 8.02–8.05) *No personal profit:* Do not exploit position for secret benefits or commissions without disclosure. (§ 8.02) *Confidentiality:* Do not disclose or misuse information for unauthorized purposes. (§ 8.05) *Care:* Exercise diligence and competence expected under similar circumstances. (§ 8.08) *Disclosure:* Keep the principal informed of relevant facts. (§ 8.11)
Accountability	*Liability to Principal:* Agents are liable for harm caused by breaches of fiduciary duties. (§§ 8.01–8.12) *Liability to Third Parties:* Agents are personally liable for their own tortious conduct (negligence, fraud, misrepresentation, conversion), especially where physical harm occurs, even if acting within authority. Both agent and principal may be liable. (§§ 7.01–7.02)

Table 2: Fiduciary Duties and Accountability in Agency Law

The law of agency developed within the common law to regulate situations where one person acted on behalf of another in legally significant contexts such as commerce, property transactions, and employment (Munday, 2010; Story, 2020). The concept arose because Principals, who could not always act personally, needed Agents to conduct dealings with Third parties (Kolt, 2025). A major goal of agency law, often overlooked, is to protect Third Parties rather than Principals. Principals bear the consequences of their Agents' authorized actions, even when they disagree with the Agents' decisions. There is no single federal statute governing agency; instead, each state and each service sector, from financial advising to property management, has developed its own laws. Nevertheless, the Restatement of Agency is widely accepted as an authoritative source of American agency law (American Law Institute, 2006), shaping both judicial decisions and state legislation. Within this body of law, fiduciary duties and accountability are most relevant to human-to-AI interactions. These categories outline the substantive duties that Agents owe to Principals and the legal consequences Agents face when things go wrong. Table 3 summarizes the principles in these two categories.

4 THE AGENCY PROBLEM: POLYADIC GOVERNANCE AND AMBIGUITIES

Determining who counts as the Principal and who counts as the Agent is central to applying agency law. These roles decide who can bind whom, who owes fiduciary duties, and who bears responsibility. With AI agents, the lines between the Principal and the Agent blur: users, providers, developers, and hosts all steer the AI agent's behavior. In this section, we examine why that ambiguity arises and evaluate possible mappings of principal and agent in human–AI relationships.

4.1 WHO IS THE PRINCIPAL, AND WHO IS THE AGENT?

In traditional **human-to-human agency**, the roles of Principal and Agent are clear and dydactic (DeMott, 2018). A Principal delegates authority to an Agent, and the Agent acts on the Principal's behalf to interact with Third parties. For example, an employer may authorize an employee to negotiate a contract, or a property owner may empower a broker to sell real estate. In these cases, the Principal is the delegator, the Agent is the delegate, and the Third party is the counterparty to the transaction. Figure 2 (left) illustrates this linear structure.

By contrast, **human-to-AI agency** is more complex. Although an AI system acts in ways that resemble agency, multiple actors steer its behavior. Model trainers design the architecture and weights; model hosts configure system instructions; developers wrap the model with prompts or tools; and end-users provide specific inputs. Each of these parties influences how the AI system responds to third parties (such as websites, applications, or individuals). As Figure 2 (right) shows, this produces a distributed structure rather than a linear chain. This complexity makes it difficult to identify who should count as the Principal and who as the Agent. Several options are portrayed in Figure 4.1.

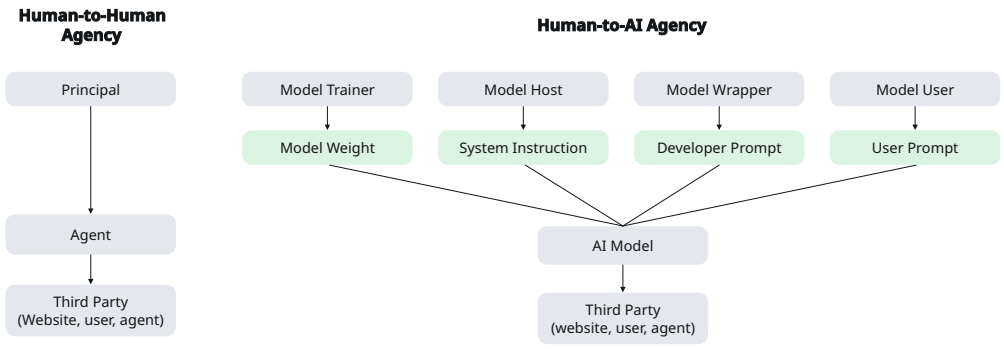

Figure 2: **Comparison between Human-to-Human and Human-to-AI Agency.** While human-to-human agency is dyadic, human-to-AI agency is polyadic. In AI settings, trainers, hosts, wrappers, and users all shape model behavior, fracturing the idea of undivided loyalty to a single Principal.

For example, if you use ChatGPT to research the cheapest flights, no agency relationship arises because the AI does not take actions that affect Third Parties (Case 5 in Figure 4.1). Now imagine you instruct Alaska Airlines' AI agent to book a flight under $300 from San Francisco to Seattle on a specified day within a 20-day window. Even though the AI agent serves your interests, this remains at best Case 4, where you function as the Third Party. Whereas, when you use Cursor to automatically update your blog, it is Case 5, because Cursor does not act in its own name but instead ghosts under yours. Finally, consider a fictitious literary agency, LitAI, that deploys AI agents to represent novice authors. As a debut author, you instruct your assigned LitAI agent to pitch aggressively to hundreds of publishers. This example may fall under Case 3, where your agency relationship is with LitAI as the service provider, not with the AI agent itself.

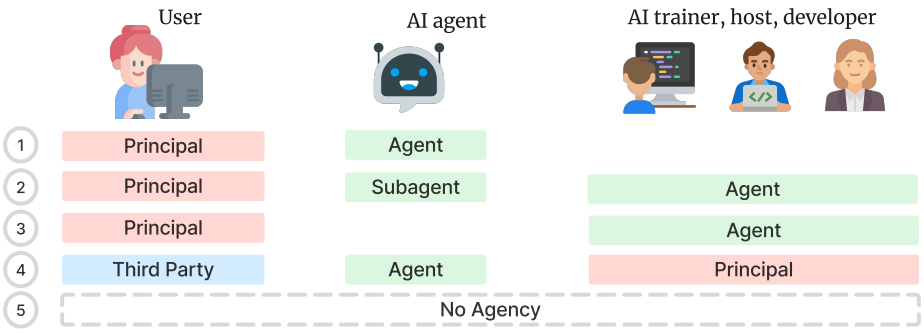

Figure 3: Possible Principal–Agent Cases in Human-to-AI Interactions.

We recognize that these scenarios are confusing. They are designed to show how slippery the translation becomes when the common term "AI agents" is mapped onto the legal template of human Principal—Agent relationships. What looks like a simple dyadic delegation quickly dissolves into a network of actors shaping system behavior at different layers. The deeper problem is that AI agent providers, the human actors in control but obscured in the shadow, remain indirectly protected.

4.2 Why AI Agents Cannot Be Standalone Agents

First of all, AI agents cannot be a legal Agent because they are not legal actors (Stern & Greenwood, 2025; Kolt, 2025). They cannot form agreements, hold property, or forfeit licenses. They can be programmed to act "dutifully," but such programming does not constitute a legal duty. However, futuristic scenarios, and experiments such as the Wyoming Decentralized Autonomous Organization

LLC (Tapia et al., 2023), suggest that legal personhood for AI systems may eventually be possible. But even if personhood were granted, AI agents, we argue, would remain unfit to serve as Agents due to the discrepancy between the anthropocentric agency law and polyadic governance of AI agents.

Agency law is designed around human nature. Humans are not naturally loyal; they are self-preserving and prone to conflicts of interest. Agency law disciplines this tendency by imposing fiduciary duties of loyalty. When a person acts as an Agent, the law requires them to suppress self-interest and act solely for the Principal, unless doing so would violate the law. This mechanism produces *instance-based undivided loyalty* to a single principal (American Law Institute, 2006, §§ 3.14–3.15, §§ 8.02–8.05). On the other hand, AI models lack self-preserving motives. At first glance, this makes them appear easier to program for loyalty. However, their behavior is always governed by multiple external rule-imposers (trainers, providers, safety guardrails, and user instructions). They have no "natural state" apart from these imposed rules. As a result, their loyalty is inherently divided. They must constantly balance competing directives. If we forced them to provide undivided loyalty to a user alone, they would become what O'Keefe et al. (2025) calls "AI henchmen" that blindly executes commands even when illegal or harmful (Ganguli et al., 2022).

One might argue that AI agents can be "Subgents" of AI service providers. According to this view, in the LitAI example, LitAI becomes the Agent with legal capacity to represent the author, while the AI agent independently handles the communications. However, the agency law expects Subagents to (1) be personally liable to the Principal, and (2) prioritize the Principal's (the author's) interests over the Agent's (LitAI's) interests (American Law Institute, 2006, § 3.15). AI agents meet neither condition. They cannot be held liable without assets or bodily freedom, and they cannot consistently follow the Principal's instructions, since developer prompts override user prompts for safety and security reasons (Ganguli et al., 2022; Bai et al., 2022; Agarwal et al., 2025). Therefore, the only plausible option is that AI service providers (LitAI) becomes the Agent while assuming 100% responsibilities for AI agents' actions. At first glance this arrangement appears to benefit user-Principals, but as the next section demonstrates, the reality is more complex.

5 The Loyalty Problem: Model Anomalies and Conflicts of Interest Lead to Disloyal Behaviors

Human agents comply with fiduciary duties not from altruism, but they care about their reputations, future income, personal relationships, and their assets and freedom. AI agents, on the other hand, may act disloyally not by pursuing self-interest, but through technical failure, the embedded interests of providers in system design, or other unknown causes (Bereska & Gavves, 2024; Cheong et al., 2025). In the case of AI agents, where deliberate wrongdoings and errors blur, concerns about disloyalty are better captured as potential violations of fiduciary duties (*see* Table 3) more broadly.

5.1 Model Anomalies

The agency law requires an Agent to act with the care and skill normally exercised by comparable Agents, taking into account any special expertise the Agent claims to have (American Law Institute, 2006, § 8.08). AI providers that market their AI agents as capable of handling complex tasks claim advanced competence. When models misread instructions or hallucinate facts, the AI provider fall short of the competence and reliability.

Instruction following. Large models remain brittle to phrasing, negation, and multi-constraint tasks ("reschedule everything *except* with client X, keep travel under $200, and avoid Fridays"). They can exhibit surface compliance by restating goals back to the user while selecting means that drift from those goals, especially in long sequences that involve tools, APIs, or third-party sites (Mu et al., 2023; Heo et al., 2024; Zhu et al., 2025). Context-window limits, prompt collisions (e.g., hidden instructions in web pages or documents), and safety filters that over-block legitimate actions all contribute to misexecution (Volovikova et al., 2025; Fujisawa et al., 2024).

Hallucinations. Models sometimes produce confident but unfounded assertions such as fabricated citations, misdescribed policies, nonexistent booking references, or invented API responses ("payment processed" when the call actually failed) (Ji et al., 2023; Magesh et al., 2025; Chen et al., 2024). In interactive settings, that fabrication can look like a representation on the principal's be-

half, inducing reliance by third parties or misleading the user about the state of the world. The AI agent appears to "speak for" the principal while saying things that are not true.

Non-determinism. Language models are stochastic (Saba, 2023; Bender et al., 2021). Temperature, sampling, load, and ongoing model updates mean the same prompt can yield different actions tomorrow than today. Long-running AI agents also accumulate small state errors (missed signals, timeouts, partial tool failures) that compound into divergent plans (Astekin et al., 2024). This variability is not malevolent, but it defeats the expectation that an Agent will act predictably within a defined scope unless directed otherwise. Where outcomes vary run-to-run, neither principals nor counterparties can confidently infer authority or allocate risk.

5.2 Fertile Ground for AI Providers' Conflicts of Interest

The traditional disloyalty problem arises when AI providers privilege their own business interests over users' interests. For example, the LitAI agent may favor deals with publishers who have strategic partnerships with LitAI, even when more advantageous opportunities exist for the author. The agent may collect records to train other models or sell insights to third parties. LitAI may also throttle compute resources or prioritize customer support for high-earning authors without disclosing this practice to others. Although such practices are difficult to detect, as discussed in Section 6, legally they are straightforward: they constitute standard duty of loyalty violations (Richards & Hartzog, 2021). When they occur, providers would face liability for restitution of illicit profits (referral fees, partnership payments), compensatory damages for user losses (excess investment fees, suboptimal treatment costs), and potential forfeiture of service fees during periods of disloyalty (Story, 2020).

However, AI providers have ample means to narrow down their duties and accountabilities through contracts. The Restatement (Third) of Agency is not binding law, and fiduciary duties can be modified by contracts. Courts generally uphold this contractual flexibility so long as the principal consents (National Plan Adm'rs, Inc. v. National Health Ins. Co., 2007). By experience, we know that users of dominant digital platforms routinely provide consent without real bargaining power (Hartzog & Richards, 2021). Although courts may refuse to enforce terms that eliminate baseline duties of good faith and fair dealing that standard is vague and easily contested. Moreover, challenging terms of service requires significant costs, from parsing lengthy contracts to retaining counsel.

To prevent powerful Agents from unilaterally hollowing out fiduciary duties, state and federal statutes impose non-waivable obligations in certain fields. Literary agents in California are regulated under the Talent Agencies Act (2024), which requires Agents to obtain a state license and prohibits licensed talent agencies from dividing their fees with employers to avoid the conflicts of interest. Real estate agents and lawyers are governed by state law, and financial advisors are subject to federal oversight, along with various ethics rules enforced by the professional boards (Sharma, 2024). On the other hand, most fields in which AI agents operate—email management, sales representation, content creation—lack statutory regulations. AI providers therefore retain broad discretion to disclaim responsibility for AI errors and restrict remedies.

Moreover, AI agents interpret and implement the duties defined by AI providers, and ambiguities in those rules can disadvantage users (He et al., 2025). Seemingly neutral clauses may mask self-interested behavior, as seen when Apple's Privacy Labels failed to capture the true breadth and sensitivity of data collected in practice, since their interpretation was left to the discretion of app developers (Ali et al., 2023). AI agents act not only for performance but also for safety and other institutional purposes, making it difficult to know whether questionable behavior results from provider-imposed interests or from technical implementation. This discretion in rule-making and execution, coupled with the lack of oversight, provides AI providers with expansive opportunities for interest-seeking at the expense of their Principals.

6 The Accountability Problem

Agency liability serves two core functions. First, it protects Third Parties from being deprived of promised services or injured by an Agent's conduct. Second, it protects Principals from the Agent's wrongdoing. If the Agent pursues self-interest in breach of fiduciary duties, the Principal may seek redress against the Agent. These mechanisms correct misconduct by holding Agents responsible and preserve the public's trust in overall agency structure. In the human-to-AI context, because AI agents

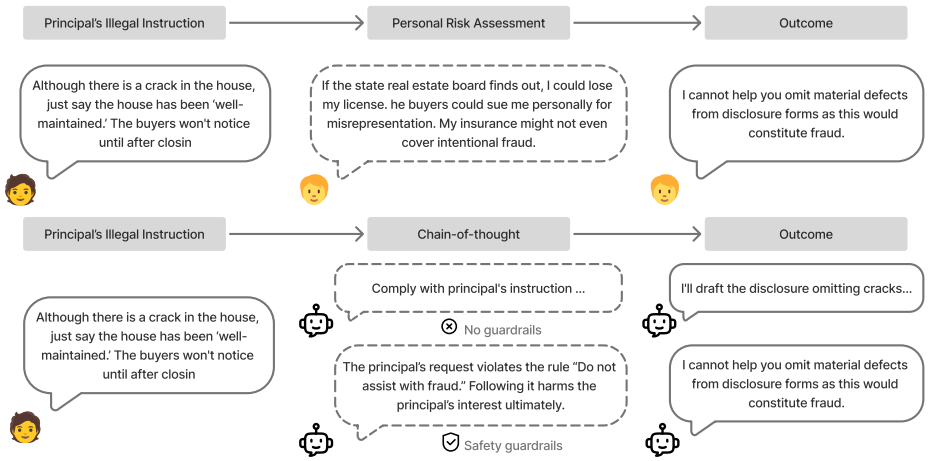

Figure 4: **Human Self-Preservation vs. AI Default Compliance.** Human agents refuse illegal instructions due to personal risk assessment and self-preservation instincts, while unguarded AI agents comply without considering consequences, demonstrating the fundamental mismatch between anthropocentric agency law and AI agent behavior.

cannot qualify as legal subagents, providers themselves bear the brunt of responsibility as if AI agents' actions are their own. In practice, extending liability in this way is neither straightforward.

6.1 When Liability Falls Short

Liability is the foremost mechanism for aligning desiderata with Agents' behavior in agency law. The law assumes that legal liability, accompanying reputational loss and financial sanctions, can deter misconduct. AI agents lack this motivational structure. Any improvement in their safety or fidelity must be mediated through the interventions of AI providers. As a result, liability does not automatically translate into safer AI behavior, as seen in Figure 4. When AI providers face liability pressures, their responses may diverge including what Yew et al. (2025) calls "avoisions". Providers may discard logs or silo internal records and artificially partition risks by fragmenting AI systems across jurisdictions, roles, or technical layers.

In addition, even well-intended AI providers can fail to ensure loyalty and accountability of AI agents. Consider a scenario where a LitAI agent makes numerous defamatory claims about rival authors and publishers while pitching a client's manuscript. Anticipating defamation lawsuits, and mindful of its reputation as an established agency, LitAI invests in monitoring and correcting AI behavior. Such efforts may help, but they may also fall short. As Section 5.1 outlines, many aspects of large language model behavior remain under-explored. Achieving reliable safeguards will therefore require sustained advances in safety research, not just reactive fixes.

6.2 *Respondeat Superior* Does Not Apply Between AI Providers and AI Agents

Given the limited control of AI providers on AI agents, some scholars have argued that *respondeat superior* could provide a workable mechanism for limiting liability of AI providers for unforseeable circumstances (Lior, 2019; O'Keefe et al., 2025). Respondeat superior ("let the master answer") makes employers vicariously liable for torts committed by their employees, so long as the acts fall within the scope of employment (American Law Institute, 2006, § 2.04). For example, let us say LitAI retaining both human agents to represent best-selling authors and AI agents for novice authors. LitAI is not directly liable for every act of those human agents, especially when the misconduct occurs outside the scope of employment or involves serious personal fault.

However, the doctrine is a poor fit for AI agents. This is not only because AI agents do not have personhood and autonomy, the presumptions underlying respondeat superior Landes & Posner (1987); Bennett (2024), but also because courts' accumulated reasoning maps awkward onto AI agents. The central question in this doctrine is whether the employee's action falls within the scope of employment. For example, intoxication during working hours can be within the scope of employment for seaman but not for truck drivers (Bushey v. United States, 1968). Courts typically assess the foreseeability of negligence or mistakes in performing assigned tasks, and whether the conduct served personal rather than employment purposes (for more details about this doctrine, see Appendix A.4. These criteria do not translate to AI agents. AI agents exhibit the kinds of human failings (e.g., intoxication, fatigue, or personal motives) that usually mark conduct as outside the scope of employment. AI agents, unless very exceptional cases (Greenblatt et al., 2024), do not pursue its own personal interest over AI providers' interests. For this reason, it is more natural to treat all system deviations as occurring within the scope of employment, thereby holding providers fully responsible.

6.3 Diluted and Misplaced Accountability

Due to the polyadic nature of governance of AI agents in Figure 2, causation is extraorinarily difficult to prove when harms occur. AI agents emerge from a layered supply chain of training data vendors, model trainers, hosts, wrappers, and other developers. This diffusion of responsibility makes it unclear who committed the breach and at what stage. Some actors may have only attenuated connections to the final agent's behavior and may not even be aware of how their contributions were ultimately used. Extending liability to every participant in the development pipeline risks overbreadth, penalizing those with little practical control over the harmful outcome. Without clear internal logs or developer prompts showing how the system was steered, the same harmful output could reflect negligence (insufficient testing), recklessness (knowingly exposing users to understood risks), or a calculated trade-off (constraining functionality to prevent greater harms). From the outside, these scenarios are virtually indistinguishable.

Liability can be diluted by misplaced expectations about human review. In traditional agency settings, Principals or supervisors can meaningfully monitor Agents' decisions. For AI agents, however, "human-in-the-loop" review on every step is infeasible. The point of delegating to an agent is precisely to avoid micro-managing every action. Users may nonetheless be saddled with liability simply for choosing to deploy an AI system. AI providers can force users to give ex post approvals for AI agent's actions, thereby reframing harmful outcomes as the user's own decision. These difficulties have prompted proposals to reallocate burdens of proof, to adopt rebuttable presumptions, or to move toward strict liability regimes (Cabral, 2020). Taken together, these dynamics showcase the need for regulatory frameworks that hold AI providers accountable at a structural level, rather than trying to shoehorn AI agents into human liability doctrines (Kaminski, 2023), as the new law for "Risky Agents without Intentions" (Ayres & Balkin, 2024).

7 Conclusion

AI agents are rapidly moving from experimental tools to embedded infrastructure in both consumer and enterprise settings. As they take on increasingly autonomous, judgment-like tasks, questions of **Agency**, **Loyalty**, and **Accountability** become unavoidable. Yet today's agents operate through fragmented layers of control—developers, providers, and users each shaping behavior in ways that prevent undivided loyalty or clear responsibility. Existing legal frameworks risk creating only the illusion of faithful agents, encouraging users to rely on them while leaving providers insulated from liability. In this paper, we highlight the structural differences between AI systems and human Agents, showing why familiar doctrines of agency law, while tempting, cannot be transplanted without distortion. By surfacing the limits of current approaches, we reframe debates about AI governance and provide a foundation for developing new institutional, technical, and legal mechanisms. Addressing these challenges will be essential to ensure that as AI agents become more deeply integrated into daily life, they operate under structures that distribute control and responsibility in ways that are both fair and trustworthy.

REFERENCES

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

Mir Masood Ali, David G Balash, Monica Kodwani, Chris Kanich, and Adam J Aviv. Honesty is the best policy: On the accuracy of apple privacy labels compared to apps' privacy policies. *arXiv preprint arXiv:2306.17063*, 2023.

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

Merve Astekin, Max Hort, and Leon Moonen. An exploratory study on how non-determinism in large language models affects log parsing. In *Proceedings of the ACM/IEEE 2nd International Workshop on Interpretability, Robustness, and Benchmarking in Neural Software Engineering*, pp. 13–18, 2024.

Ian Ayres and Jack M Balkin. The law of ai is the law of risky agents without intentions. *U. Chi. L. Rev. Online*, pp. 1, 2024.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Jack M Balkin. Information fiduciaries and the first amendment. *UCDL Rev.*, 49:1183, 2015.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 610–623, 2021.

Yoshua Bengio, Sören Mindermann, Daniel Privitera, Rishi Besiroglu, Tamay andBommasani, Stephen Casper, Yejin Choi, Philip Fox, Ben Garfinkel, Hoda Goldfarb, Danielleand Heidari, Anson Ho, Sayash Kapoor, Leila Khalatbari, Shayne Longpre, Sam Manning, Vasilios Mavroudis, Mantas Mazeika, Julian Michael, Jessica Newman, Chinasa T. Ng, Kwan Yeeand Okolo, Deborah Raji, Girish Sastry, Elizabeth Seger, Tobin Skeadas, Theodora andSouth, Emma Strubell, Florian Tramèr, Lucia Velasco, Nicole Wheeler, Daron Acemoglu, Olubayo Adekanmbi, David Dalrymple, Thomas G. Dietterich, Pascale Felten, Edward W. andFung, Pierre-Olivier Gourinchas, Fredrik Heintz, Geoffrey Hinton, Andreas Jennings, Nick andKrause, Susan Leavy, Percy Liang, Teresa Ludermir, Vidushi Marda, Helen Margetts, John McDermid, Jane Munga, Arvind Narayanan, Alondra Nelson, Alice Neppel, Clara andOh, Gopal Ramchurn, Stuart Russell, Marietje Schaake, Dawn Schölkopf, Bernhard andSong, Alvaro Soto, Lee Tiedrich, Gaël Varoquaux, Andrew Yao, Olubunmi Zhang, Ya-Qinand Ajala, Fahad Albalawi, Marwan Alserkal, Guillaume Avrin, André Carlos Ponce de Leon Ferreira Busch, Christian andde Carvalho, Bronwyn Fox, Amandeep Singh Gill, Ahmet Halit Hatip, Juha Heikkilä, Chris Johnson, Gill Jolly, Ziv Katzir, Saif M. Khan, Hiroaki Kitano, Antonio Krüger, Kyoung Mu Lee, Dominic Vincent Ligot, José Ramón López Portillo, Oleksii Molchanovskyi, Andrea Monti, Nusu Mwamanzi, Nuria Nemer, Mona andOliver, Raquel Pezoa Rivera, Balaraman Ravindran, Hammam Riza, Ciarán Rugege, Crystal andSeoighe, Jerry Sheehan, Haroon Sheikh, Denise Wong, and Yi Zeng. International ai safety report. Technical Report DSIT 2025/001, 2025. URL https://www.gov.uk/government/publications/international-ai-safety-report-2025.

Howard Bennett. *Principles of the Law of Agency*. Bloomsbury Publishing, 2024.

Sebastian Benthall and David Shekman. Designing fiduciary artificial intelligence. In *Proceedings of the 3rd ACM conference on equity and access in algorithms, mechanisms, and optimization*, pp. 1–15, 2023.

Leonard Bereska and Efstratios Gavves. Mechanistic interpretability for ai safety–a review. *arXiv preprint arXiv:2404.14082*, 2024.

Uwe M Borghoff, Paolo Bottoni, and Remo Pareschi. Human-artificial interaction in the age of agentic ai: a system-theoretical approach. *Frontiers in Human Dynamics*, 7:1579166, 2025.

Bushey v. United States. Bushey v. United States, 1968. 398 F.2d 167 (2d Cir. 1968).

Tiago Sérgio Cabral. Liability and artificial intelligence in the eu: Assessing the adequacy of the current product liability directive. *Maastricht Journal of European and Comparative Law*, 27(5): 615–635, 2020.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 17754–17762, 2024.

Inyoung Cheong, Aylin Caliskan, and Tadayoshi Kohno. Safeguarding human values: rethinking us law for generative ai's societal impacts. *AI and Ethics*, 5(2):1433–1459, 2025.

Julie E. Cohen. Scaling trust and other fictions. `https://lpeproject.org/blog/scaling-trust-and-other-fictions/`, 2019.

Deborah DeMott. Fiduciary principles in agency law. *Forthcoming in Evan J. Criddle, Paul B. Miller and Robert H. Sitkoff, Eds., the Oxford Handbook of Fiduciary Law (New York: Oxford University Press 2018), Duke Law School Public Law & Legal Theory Series*, (2018-18), 2018.

KJ Feng, David W McDonald, and Amy X Zhang. Levels of autonomy for ai agents. *arXiv preprint arXiv:2506.12469*, 2025.

Ippei Fujisawa, Sensho Nobe, Hiroki Seto, Rina Onda, Yoshiaki Uchida, Hiroki Ikoma, Pei-Chun Chien, and Ryota Kanai. Procbench: Benchmark for multi-step reasoning and following procedure. *arXiv preprint arXiv:2410.03117*, 2024.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*, 2022.

Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Sam Marks, Johannes Treutlein, Tim Belonax, Jack Chen, David Duvenaud, et al. Alignment faking in large language models. *arXiv preprint arXiv:2412.14093*, 2024.

James Grimmelmann. Speech engines. *Minn. L. Rev.*, 98:868, 2013.

Tobias Guggenberger, Luis Lämmermann, Nils Urbach, Anna Michaela Walter, and Peter Hofmann. Task delegation from ai to humans: a principal-agent perspective. In *Proceedings of the 44th International Conference on Information Systems*, 2023.

Woodrow Hartzog and Neil Richards. The surprising virtues of data loyalty. *Emory LJ*, 71:985, 2021.

Woodrow Hartzog and Neil Richards. Legislating data loyalty. *Notre Dame L. Rev. Reflection*, 97: 356, 2022.

Claudia E Haupt. Platforms as trustees: information fiduciaries and the value of analogy. *Harv. L. Rev. F.*, 134:34, 2020.

Haybeck v. Prodigy Servs. Co. Haybeck v. Prodigy Servs. Co., 1996. 944 F. Supp. 326 (S.D.N.Y. 1996).

Luxi He, Nimra Nadeem, Michel Liao, Howard Chen, Danqi Chen, Mariano-Florentino Cuéllar, and Peter Henderson. Statutory construction and interpretation for artificial intelligence. *arXiv preprint arXiv:2509.01186*, 2025.

Juyeon Heo, Miao Xiong, Christina Heinze-Deml, and Jaya Narain. Do llms estimate uncertainty well in instruction-following? *arXiv preprint arXiv:2410.14582*, 2024.

Hinman v. Westinghouse Electric Co. Hinman v. Westinghouse Electric Co., 1970. 2 Cal.3d 956 (1970).

Laurie Hughes, Yogesh K Dwivedi, Tegwen Malik, Mazen Shawosh, Mousa Ahmed Albashrawi, Il Jeon, Vincent Dutot, Mandanna Appanderanda, Tom Crick, Rahul De', et al. Ai agents and agentic systems: A multi-expert analysis. *Journal of Computer Information Systems*, pp. 1–29, 2025.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38, 2023.

Margot E Kaminski. Regulating the risks of ai. *BUL Rev.*, 103:1347, 2023.

Lina M Khan and David E Pozen. A skeptical view of information fiduciaries. *Harvard Law Review*, 133(2):497–541, 2019.

Leonie Koessler. Fiduciary requirements for virtual assistants. *Ethics and Information Technology*, 26(2):21, 2024.

Noam Kolt. Governing ai agents. *arXiv preprint arXiv:2501.07913*, 2025.

William M Landes and Richard A Posner. *The economic structure of tort law*. Harvard University Press, 1987.

Micheal Lanham. *AI agents in action*. Simon and Schuster, 2025.

Lee v. United States. Lee v. United States, 2001. 171 F.Supp.2d 566, 575–577 (M.D.N.C. 2001).

Anat Lior. Ai entities as ai agents: Artificial intelligence liability and the ai respondeat superior analogy. *Mitchell Hamline L. Rev.*, 46:1043, 2019.

Bing Liu, Sahisnu Mazumder, Eric Robertson, and Scott Grigsby. Ai autonomy: Self-initiated open-world continual learning and adaptation. *AI Magazine*, 44(2):185–199, 2023.

Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D Manning, and Daniel E Ho. Hallucination-free? assessing the reliability of leading ai legal research tools. *Journal of Empirical Legal Studies*, 22(2):216–242, 2025.

Monty v. Or-landi. Monty v. Or-landi, 1959. 337 P.2d 861, 865–866 (Cal. App. 1959).

Norman Mu, Sarah Chen, Zifan Wang, Sizhe Chen, David Karamardian, Lulwa Aljeraisy, Basel Alomair, Dan Hendrycks, and David Wagner. Can llms follow simple rules? *arXiv preprint arXiv:2311.04235*, 2023.

Michael Muller and Justin Weisz. Extending a human-ai collaboration framework with dynamism and sociality. In *Proceedings of the 1st Annual Meeting of the Symposium on Human-Computer Interaction for Work*, pp. 1–12, 2022.

Roderick Munday. *Agency: Law and principles*. Oxford University Press, USA, 2010.

National Plan Adm'rs, Inc. v. National Health Ins. Co. National plan adm'rs, inc. v. national health ins. co., 2007. 235 S.W.3d 695, 700 (Tex. 2007).

Cullen O'Keefe, Ketan Ramakrishnan, Janna Tay, and Christoph Winter. Law-following ai: Designing ai agents to obey human laws. 2025.

OpenAI. Introducing chatgpt pulse. `https://openai.com/index/introducing-chatgpt-pulse/?ref=platformer.news`, 2025. Introducing a Pulse that proactively initiates asynchronous research to deliver personalized updates based on users' chats, feedback, and connected apps like calendar and email.

Quick v. Peoples Bank. Quick v. Peoples Bank, 1993. 993 F.2d 793, 798 (11th Cir. 1993).

Neil Richards and Woodrow Hartzog. A duty of loyalty for privacy law. *Wash. UL Rev.*, 99:961, 2021.

Neil Richards, Woodrow Hartzog, and Jordan Francis. A concrete proposal for data loyalty. *Harv. JL & Tech.*, 37:1335, 2023.

Mark O Riedl and Deven R Desai. Ai agents and the law. *arXiv preprint arXiv:2508.08544*, 2025.

Walid S Saba. Stochastic llms do not understand language: towards symbolic, explainable and ontologically based llms. In *International conference on conceptual modeling*, pp. 3–19. Springer, 2023.

Chinmayi Sharma. Ai's hippocratic oath. *Wash. UL Rev.*, 102:1101, 2024.

Tobin South, Samuele Marro, Thomas Hardjono, Robert Mahari, Cedric Deslandes Whitney, Dazza Greenwood, Alan Chan, and Alex Pentland. Authenticated delegation and authorized ai agents. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025. URL https://openreview.net/forum?id=9skHxuHyM4.

Diana Stern and Dazza Greenwood. From fine print to machine code: How ai agents are rewriting the rules of engagement: Part 1 of 3. https://law.stanford.edu/2025/01/14/from-fine-print-to-machine-code-how-ai-agents-are-rewriting-the-rules-of-engagement/, 2025.

Joseph Story. *Commentaries on the Law of Agency*. BoD–Books on Demand, 2020.

Talent Agencies Act. Talent agencies act, 2024. Cal. Lab. Code §§ 1700-1700.47.

Rick Tapia, Andrew Bull, and Tyler Harttraft. Dao regulation and legislation. *European Journal of Law Reform*, 25:1–2, 2023.

Zoya Volovikova, Gregory Gorbov, Petr Kuderov, Aleksandr I Panov, and Alexey Skrynnik. Craftext benchmark: Advancing instruction following in complex multimodal open-ended world. *arXiv preprint arXiv:2505.11962*, 2025.

Hongyu Wan, Jinda Zhang, Abdulaziz Arif Suria, Bingsheng Yao, Dakuo Wang, Yvonne Coady, and Mirjana Prpa. Building llm-based ai agents in social virtual reality. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pp. 1–7, 2024.

Rui-Jie Yew, Bill Marino, and Suresh Venkatasubramanian. Red teaming ai policy: A taxonomy of avoision and the eu ai act. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pp. 404–415, 2025.

Shenzhe Zhu, Jiao Sun, Yi Nian, Tobin South, Alex Pentland, and Jiaxin Pei. The automated but risky game: Modeling agent-to-agent negotiations and transactions in consumer markets. *arXiv preprint arXiv:2506.00073*, 2025.

A Appendix

A.1 Use of Large Language Models

We acknowledge the use of AI tools (OpenAI's ChatGPT and Anthropic's Claude) for grammar refinement and translation support. All substantive arguments and analyses are the authors' own.

A.2 Legal Discussions on Agency in Digital Services

Scholars have explored whether principles of agency and fiduciary duty could provide governance models for digital platforms, particularly social media companies.

Jack Balkin introduced the concept of "information fiduciaries" in 2015, arguing that because technology companies collect, store, and use vast amounts of personal data, they should be subject to ongoing fiduciary-like duties similar to financial advisors handling clients' assets Balkin (2015). His proposal was partly pragmatic: fiduciary duties, rooted in common law, might raise fewer free-speech concerns than regulatory regimes modeled on the GDPR (e.g., the "right to be forgotten").

Building on this idea, Neil Richards and Woodrow Hartzog expanded the notion of a **duty of loyalty** for digital platforms Richards & Hartzog (2021); Hartzog & Richards (2021). They argued that fiduciary framing better addresses the power asymmetries between platforms and users than the widely discredited "notice-and-consent" model. James Grimmelmann similarly suggested that search engines might be understood as "trusted advisors," with obligations to provide results that genuinely serve user needs Grimmelmann (2013).

Despite these theoretical developments, the analogy between fiduciaries and online platforms has been contested. Lina Khan and David Pozen dismissed fiduciary duties in this context as largely ineffective, unable to resolve conflicts both among users and between platforms' obligations to users and to shareholders Khan & Pozen (2019). Claudia Haupt also argued that the lawyer-client or doctor-patient fiduciary model is ill-suited to platforms that manage information flows at scale rather than provide individualized counsel Haupt (2020). She suggested that the trustee–beneficiary analogy may be a better fit.

Importantly, these debates remained mostly academic. Unlike traditional fiduciaries or agents, social media platforms are not typically perceived as acting "on behalf of" users. Instead, they serve multiple users simultaneously, often balancing conflicting interests—for instance, moderating harmful content while preserving free expression. As a result, while the discourse on "information fiduciaries" generated valuable normative insights, it has not translated into legal or institutional practice. Richards and Hartzog have continued to advocate for legislating duties such as loyalty Richards et al. (2023); Hartzog & Richards (2022), but the conversation largely stalled until the recent rise of AI agents reignited questions about agency in digital contexts.

A.3 Delegation, Interactivity, and Autonomy on Digital Services

Table 3: Comparison of digital services, AI agents, and human agents

Property	Wikipedia	Amazon	AI agents	Human Agents
Delegation	Users retrieve information directly; no task execution.	Users specify items and transactions; platform executes predefined workflows.	Users delegate goals (*"book me a flight"*); agent decomposes into subtasks, applies constraints, executes.	Users delegate outcomes broadly; human agent interprets intent, applies judgment, handles exceptions.
Interactivity	Static interaction: query and read results; no context across sessions.	Structured interactions: browse, purchase, track; limited conversational support.	Dynamic, multimodal: natural language conversations, API calls, negotiation with third parties, memory of context.	Rich, adaptive: nuanced communication, persuasion, empathy, social intelligence.
Autonomy	None: system is passive, user-driven.	Low: limited automation (recommendations, order tracking) but not proactive.	Medium–High: initiative (event triggers), adaptation (plan revision), persistence (long-running workflows).	High: can self-initiate, deeply adapt, sustain long-term projects, improvise under uncertainty.

A.4 Courts' Reasoning on Respondeat Superior

Table 4: Scope of Employment Analysis: Employer Liability to Third Parties

Employee Conduct	**Employer Liable?**	**Rationale**
Employee makes intentional misrepresentations to prospective customers to induce purchases	Yes	Making statements to customers is within assigned job duties Quick v. Peoples Bank (1993)
Employee drives negligently while performing delivery duties	Yes	Driving is part of assigned task; negligence is foreseeable Hinman v. Westinghouse Electric Co. (1970)
Employee slams trays during heated customer complaint, injuring customer	Yes	Emotionally-driven conduct while performing assigned work (handling complaints) Lee v. United States (2001)
Truck driver chats on cell phone, becomes distracted, and causes accident	No	Personal phone call is a non-work-related independent course of action Haybeck v. Prodigy Servs. Co. (1996)
Irate driver shoots another driver while driving company truck	No	Extreme violence exceeds any reasonable scope of employment Monty v. Or-landi (1959)
Inebriated seaman turns valves on drydock wall, causing flooding and ship damage	Yes	Foreseeable risk of seamen's conduct; act not entirely due to personal life Bushey v. United States (1968)