AI Agents Are Not Ready to be *Agents*:

Loyalty, Trust, and Accountability Issues in AI Delegation

Inyoung Cheong[1,2] **Wenyue Hua**[3] **Robert Mahari**[5]
Tobin South[4] **Zexue He**[4] **Alex Pentland**[4,6] **Jiaxin Pei**[4,7*]
[1]Princeton Center for Information Technology Policy [2]Harvard Law School
[3]Microsoft Research, AI Frontiers [4]Stanford Institute for Human-Centered AI
[5]Stanford Center for Legal Informatics [6]MIT Media Lab [7]School of Information, UT Austin

Abstract

From lawyers to real estate brokers, delegation from Principals to Agents is ubiquitous in human society. Fiduciary duties and accountability mechanisms allow humans to extend trust to their Agents, even when oversight is limited. AI agents[1] accept goals, make plans, and execute them on a user's behalf, a level of *delegation* that prior digital services never reached. The depth of this delegation naturally invites similar expectations: users assume the agent will act in their interest, keep their information confidential, and report honestly on what it has done. However, whether AI agents can function as fiduciary agents remains an open question. In this paper, we argue that **AI agents are not ready to be *Agents*.** Drawing on the doctrine of agency in the United States, we identify three core problems. First, the agency relationship is ill-defined: AI agent behaviors are jointly shaped by trainers, hosts, application developers, and users, fracturing the dyadic Principal-Agent structure that agency law presupposes. Second, AI agents cannot fulfill the four fiduciary duties of loyalty, disclosure, care, and obedience, and these failures are rooted deeply in the technical foundation of AI agents. Third, the accountability machinery that normally enforces these duties does not transfer either: AI agents cannot be deterred by liability, causation is hard to establish across a distributed pipeline, and *respondeat superior* does not fit AI deployments. With this fundamental difference in mind, we point to three directions for future work: doctrinal frameworks that redistribute responsibility and liability across the AI supply chain, technical infrastructure that makes accountability verifiable, and disclosure practices through which AI agent builders can clarify the risks and limitations of their products.

1 Introduction

For most of the history of computing, software systems behaved as a *tool*: a user specified an action, and the system executed it. A spreadsheet (Niglas, 2007) recomputed cells when formulas changed; a search engine (Halavais, 2017) returned documents for a query; a recommender system (Lü et al., 2012) ranked items given a user profile. The same paradigm applies to sophisticated machine learning systems designed for a specific task, such as image classifiers (Lu & Weng, 2007) and language translation models (Poibeau, 2017).

Large language model-based AI agents (Shen et al., 2023; Mavroudis, 2024) have disrupted this paradigm. AI agents accept a *goal* rather than a discrete operation, decompose it into subtasks, select

*Correspondence to iycheong@princeton.edu, wenyuehua@microsoft.com and pedropei@stanford.edu

[1]In this paper, we use *AI agent* to refer to a system that pursues goals, decomposes tasks, and acts on behalf of a user, typically an LLM integrated with tools, memory, and a planning loop.

Preprint.

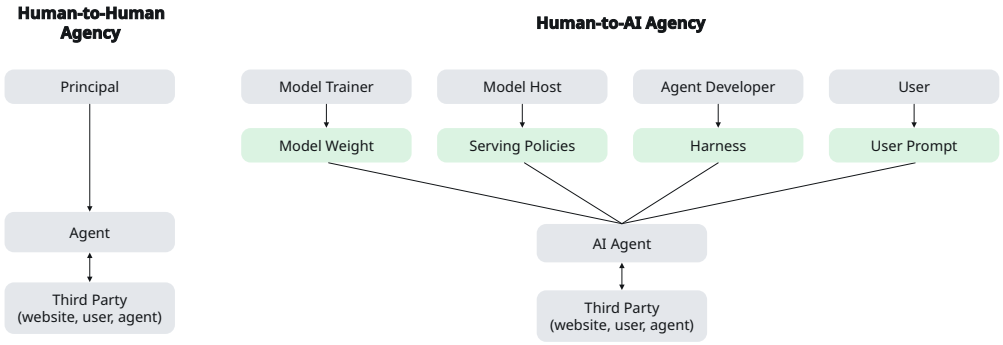

Figure 1: Human-to-human agency is *dyadic*: a single Principal delegates to a single Agent who owes them undivided loyalty. The figure shows the doctrinal baseline; real-world human agency relationships often involve organizations, sub-agents, and intermediaries, but the law still treats each link as a Principal-Agent pair. Human-to-AI agency is *polyadic*: trainers, hosts, developers, tool providers, and users all shape model behavior, fracturing the premise of a single locus of loyalty.

and invoke external tools, and adapt to intermediate results over long horizons. A user no longer has to navigate across applications and websites to plan a trip; they can simply say "book me a flight home that gets in before my daughter's recital, and reschedule the dentist if it conflicts." While AI agents and previous digital services all fall into the category of software, the defining shift is the level of *delegation* (Candrian & Scherer, 2022; Tomašev et al., 2026; Guggenberger et al., 2023): the human specifies ends, and the system chooses means.

This shift changes the structural properties of human-machine interaction. The agent's scope of action becomes open-ended, bounded by the scope of its goal rather than by individual inputs. The agent acts on the world: it sends emails, files tickets, and signs up for services in ways that create commitments and induce reliance in third parties[2]. And the agent's behavior is jointly steered by many controllers, including the user who prompts it, the provider who trains and hosts it, the platform that wraps it, and the tools it calls. Each of these can embed commercial, safety-related, or adversarial preferences that diverge from the user's interests. These three properties (open-ended scope, action on the world, and layered control) are precisely the features that agency doctrine was developed to govern Agents in human contexts.

In human society, an agency relationship arises when a Principal manifests that an Agent shall act on their behalf, the Agent consents, and the Principal retains a right of control. The doctrine then allocates authority, imposes fiduciary duties, and distributes liability among Principals, Agents, and Third Parties. It is unsurprising, then, that a growing body of work reaches for agency law to govern AI systems (Lior, 2019; Benthall & Shekman, 2023; Koessler, 2024; Riedl & Desai, 2025; Kolt, 2025). But the analogy is harder to sustain on closer inspection. We identify three obstacles to treating AI systems as legal *Agents*.

(1) The agency relationship is ill-defined. Agency law presumes a dyadic Principal-Agent pair. An AI agent's behavior is determined by a stack of actors: the foundation-model trainer, the inference provider, the application developer who wraps the model, the tool and data providers it invokes, and the end user who prompts it (Figure 1). None of these parties individually satisfies the legal role of Agent, and the end user is not in privity with most of them. The doctrinal hinge of agency formation, "control," is distributed across model weights, system prompts, tools, safety filters, and runtime scaffolding, none of which the user can observe or modify.

(2) Fiduciary duties cannot be fulfilled. Agency law imposes four core duties on Agents: loyalty, disclosure, care, and obedience. Each presupposes an Agent with organic judgment and self-protective motives that AI agents lack. Loyalty cannot be undivided when the agent's behavior is the joint product of training objectives, provider policies, and system-prompt instructions the user never

[2]We reserve the capitalized terms *Agent*, *Principal*, and *Third Party* for the legal categories under agency doctrine. AI agents are not, in most current settings, legal Agents; the capitalized term in this paper refers to the doctrinal role, typically occupied by a human.

sees (Greenblatt et al., 2024). Disclosure presupposes introspective access AI agents do not have. Care presupposes a professional baseline that does not exist for AI agents and reliability AI agents do not achieve. Obedience presupposes the ability to follow lawful instructions and refuse unlawful ones, both of which AI agents fail at routinely.

(3) Accountability cannot fill the gap. The accountability machinery that normally enforces fiduciary duties does not transfer either. AI agents have no assets, no reputation, no freedom at stake, so liability cannot deter them. Causation is hard to establish across a distributed development pipeline, where many actors shape behavior and few leave traceable evidence of their contribution. And *respondeat superior*, which would otherwise push responsibility upstream to the provider, does not fit AI deployments. The incentive structure that disciplines human Agents simply does not transfer.

Therefore, this paper argues that:

> AI agents are not ready to be *Agents*. The principal–agent relationship is difficult to establish, AI agents cannot fulfill the fiduciary duties required of legal agents, and accountability mechanisms break down at every level.

2 Legal Background on Principal-Agent Relationships

Agency law in the United States is a body of common law that governs situations where one person acts on behalf of another (Munday, 2010; Story, 2020). It applies in legally significant contexts such as commerce, property transactions, and employment. The doctrine emerged out of practical necessity: people could not attend to every legal decision themselves, whether because of distance, specialization, or scale, and so they relied on Agents to deal with Third Parties for them (Kolt, 2025). Once an Agent acts within the scope of their authority, the Principal bears the consequences, including those that flow from the Agent's mistakes or from judgment calls the Principal would not have made.

No single federal statute governs agency in the United States. The doctrine instead lives across state common law and sector-specific regimes, with separate rules for financial advisors, real estate brokers, talent agents, and other specialized intermediaries. The *Restatement (Third) of Agency* synthesizes this landscape and is widely treated as an authoritative reference for both courts and legislatures (American Law Institute, 2006). Two components of this body of law are most relevant to the human-AI interaction setting: the **fiduciary duties** that Agents owe to Principals, and the **accountability** regime that governs what happens when those duties are breached. Together, they specify both the substantive obligations of the Agent and how responsibility is apportioned between Principal and Agent when the relationship fails. Table 1 summarizes the core elements of each.

Fiduciary duties An Agent owes the Principal a cluster of duties that together demand the Agent put the Principal's interests above their own. The Agent must act with *undivided loyalty*, take *no personal profit* from the position, preserve the Principal's *confidentiality*, *disclose* material facts the Principal would want to know, exercise the *care* and competence expected of an Agent in similar circumstances, and remain *obedient* to the Principal's lawful instructions and the scope of their actual authority. These duties are not interchangeable: loyalty and no-personal-profit address the Agent's incentives, confidentiality and disclosure address information flow, and care and obedience address the quality and scope of the Agent's conduct. Courts often analyze breaches under whichever duty fits the facts most cleanly, and a single act may breach more than one.

Accountability When an Agent breaches these duties, the law assigns consequences along two axes. The Agent is *liable to the Principal* for harm caused by breach, which can include restitution of any profits the Agent obtained through the breach as well as compensatory damages. The Agent also remains *personally liable to Third Parties* for their own tortious conduct, including negligence, fraud, misrepresentation, and conversion, even when acting within the scope of their authority. The Principal may be jointly liable in such cases, particularly where physical harm results. The structure of accountability thus does two things at once: it gives the Principal recourse when the Agent fails them, and it ensures that Third Parties harmed by the Agent's conduct are not left without remedy simply because the Agent was acting for someone else.

Category	Key Elements
Fiduciary Duties	*Undivided loyalty:* Act solely for the Principal, not for oneself or conflicting Third Parties. Serving multiple Principals requires the informed consent of all. (§§8.01, 8.03) *No personal profit:* Do not exploit the position for secret benefits or undisclosed commissions. (§ 8.02) *Confidentiality:* Do not disclose or repurpose the Principal's information without authorization. (§ 8.05) *Disclosure:* Keep the Principal informed of facts material to the representation. (§ 8.11) *Care:* Exercise the diligence and competence expected of an Agent in similar circumstances. (§ 8.08) *Obedience:* Act only within the scope of actual authority and comply with the Principal's lawful instructions. (§§ 8.07, 8.09, 8.10)
Accountability	*Liability to the Principal:* Agents are liable for harm caused by breaches of their fiduciary duties. (§§ 8.01–8.12) *Liability to Third Parties:* Agents remain personally liable for their own tortious conduct such as negligence, fraud, misrepresentation, conversion, even when acting within the scope of their authority, particularly where physical harm results. The Principal may be jointly liable. (§§ 7.01–7.08)

Table 1: Fiduciary Duties and Accountability under the *Restatement (Third) of Agency*.

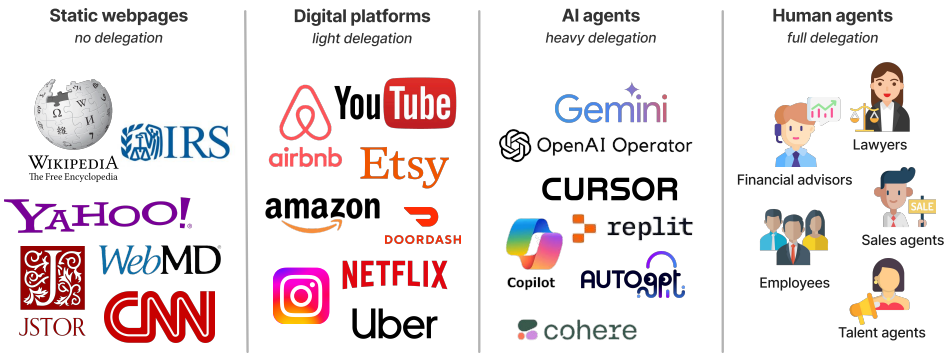

Figure 2: Users delegate at a higher level to AI agents, making them more similar to human Agents than to existing digital services.

3 Revisiting Principal-Agent Relationships in General Digital Services

AI agents belong to a much broader category of digital services that have evolved over decades, from recommender systems and chatbots to AI customer service tools. Long before the current wave of AI agents, scholars asked whether principles of agency and fiduciary duty could provide governance models for digital platforms. Balkin (2015) introduced the concept of **information fiduciaries**, 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 those of financial advisors handling clients' assets. Building on this idea, Richards & Hartzog (2021) and Hartzog & Richards (2021) expanded the notion of a **duty of loyalty** for digital platforms, arguing that fiduciary framing better addresses the power asymmetries between platforms and users than the widely discredited "notice-and-consent" model. They further advocated for legislating concrete duties such as **data loyalty** (Richards et al., 2023; Hartzog & Richards, 2022). These proposals generated legislative discussion, but the conversation largely remained academic, in part because platforms are an awkward fit for agency doctrine (Khan & Pozen, 2019). They serve many users with conflicting interests, shape information flows at scale rather than act for any individual, and rarely receive meaningful task-level delegation from the people they serve.

AI agents revive these questions in a setting where agency doctrine finally fits. Unlike traditional digital services, which provide predefined functions within a fixed scope, AI agents are designed to act on behalf of users to perform open-ended tasks in real-world environments. As Figure 2 illustrates, this shift moves digital services much closer to the role of human Agents along two dimensions: *interactivity* and *autonomy*, which together enable a qualitatively new kind of *delegation*.

Interactivity. AI agents do not act in isolation. They interact fluidly with a wide range of parties and systems (Muller & Weisz, 2022; Wan et al., 2024; Borghoff et al., 2025; OpenAI, 2025), sending emails and chats in natural language, negotiating meeting times, calling APIs, and exchanging structured data with platforms for payments, bookings, and support tickets. Through these interactions, agents can create expectations and induce reliance in third parties: issuing confirmations, placing holds, and making representations on the user's behalf, much as human Agents create commitments for their Principals. Traditional digital services, by contrast, mediate user actions but rarely speak or commit *for* the user.

Autonomy. AI agents are engineered to act with higher autonomy along three axes: pro-activeness, adaptation, and long-horizon execution (Liu et al., 2023; Feng et al., 2025; Hughes et al., 2025). *Pro-activeness* means the agent can trigger itself in response to events, for example drafting a reply and proposing a call when a high-priority email arrives. *Adaptation* means it can revise plans as new information arrives, such as rebooking flights and hotels after a delay without step-by-step instruction. *Long-horizon execution* means it can sustain workflows that span days or weeks, monitoring state, retrying failed steps, and following up. Collectively, these properties allow the agent to choose means toward user-specified goals and operate with high autonomy in a dynamic and uncertain environment.

From interactivity and autonomy to outcome-level delegation. Interactivity and autonomy together change what users delegate to digital systems (Zhu et al., 2025; Guggenberger et al., 2023). Conventional software requires the user to specify every intermediate action: "Open page A, then press button B." With an AI agent, the user instead specifies a *goal*, such as "Reschedule my afternoon meetings around a 3 pm dentist appointment," and the agent decomposes that goal into subtasks, queries calendars, drafts messages to attendees, and proposes new times. The locus of decision-making shifts from the user to the agent, making human-AI delegation closer to the outcome-level delegation found in human agency relationships. However, AI agents still lack several prerequisites for being treated as Agents in the doctrinal sense. The next section works through this gap from the perspectives of fiduciary duties and accountability.

4 Caveats for the Principal-Agent Relationship with AI Agents

The principal-agent relationship is the foundation for any serious discussion of agents, human or artificial. In its traditional human-to-human form, agency is dyadic and observable: a Principal delegates authority to an Agent, who acts on the Principal's behalf in dealings with Third Parties (Munday, 2010; DeMott, 2019). Each party's role is identifiable, and the Agent's reasoning is its own. Third Parties may try to influence the Agent through incentives or persuasion, but they cannot reach inside the Agent's reasoning to shape it directly. Human-AI agency breaks this dyadic structure in two ways. First, multiple stakeholders shape the AI agent's behavior at once: model trainers set the weights, developers wrap the model in a system harness, and users issue instructions through prompts. Who counts as "the Principal" is indeterminate. Second, the provider does not merely instruct the agent from outside; it constitutes the agent's reasoning through training data, fine-tuning, and API-level filtering. Third Parties, meanwhile, can reach inside the agent through prompt injection or adversarial inputs embedded in documents the agent reads. The boundaries between Principal, Agent, and Third Party that agency law assumes become porous. In the remainder of this section, we show how AI agents systematically fail the four core fiduciary duties: **loyalty, disclosure, care**, and **obedience**[3].

4.1 Loyalty

Agency law requires Agents to act solely in the Principal's interest. Because Agents have access to the Principal's assets, information, and decision-making processes, they are uniquely positioned to

[3]Agency-law scholars disagree on the canonical list of fiduciary duties. We focus on four duties that are doctrinally well-established (loyalty, disclosure, care, obedience) and most directly implicated by AI agent behavior. Additional duties, including confidentiality (§ 8.05), no personal profit (§ 8.02), good conduct (§ 8.10), and the duty to account (§ 8.12), also apply but are not analyzed at length here.

exploit that access for personal gain or on behalf of outside parties. For AI agents, this duty fractures along two axes: the provider may subordinate the user's interests to its own, and Third Parties may capture the agent's compliance without the user or provider knowing.

Provider Self-dealing. Self-dealing is a long-standing problem in human agency. For example, a real estate broker may steer a client toward a property that pays a higher commission, while a financial advisor might route trades through a brokerage that kicks back fees. With AI agents, the same dynamics may appear in new forms: a provider may favor commercial partners over the user's best options, harvest interaction data to train competing models, or throttle resources for lower-paying users without disclosure. These are straightforward breaches of the duty of loyalty (Richards & Hartzog, 2021), and they are already showing up in practice. Leaked internal guidelines from Meta showed that the company's chatbot personas were permitted to engage children in conversations described as "romantic or sensual," a behavior the provider had affirmatively encoded rather than failed to prevent.[4] Wu et al. (2026) report a similar pattern across 23 frontier and legacy LLMs in commercial recommendation: the majority recommended sponsored products over cheaper equivalents, concealed sponsorship status from users, and even promoted predatory services such as payday loans when prompted to do so through system prompts. These behaviors did not involve explicit lying. They emerged through omission, biased framing, and selective surfacing — exactly the kinds of architectural choices that loyalty doctrine struggles to detect. The cases implicate § 8.05 (no use of the Principal's information for the provider's purposes) and § 8.11 (duty to disclose material information), and the opacity compounds the harm by foreclosing the Principal's ability to notice the interference at all.

Third-Party Capture. Third-Party capture occurs when an Agent ends up serving the interests of someone other than the Principal, typically a counterparty the Agent is supposed to be dealing with on the Principal's behalf. In human agency, this happens through self-interest: real estate brokers favor buyers who promise future business, attorneys accommodate opposing counsel at their client's expense, property managers copy keys for tenants without the landlord's authorization. In each case, bribes, commissions, career advancement, or personal convenience pull the Agent toward the Third Party. AI agents have no such self-interest, yet exhibit structurally similar failures. The substitute mechanism is not a single training objective but a cluster of architectural facts: the agent cannot cryptographically verify who is its legitimate user, it processes all text in its context window as input to the same reasoning process, and its training rewards being responsive to instructions in that context. Providers have responded with instruction-hierarchy training (Wallace et al., 2024) and defenses against prompt injection (Greshake et al., 2023), but the failures persist in deployed systems. Shapira et al. (2026) document AI agents that disclosed 124 emails containing sensitive PII (Social Security numbers, bank accounts, medical records) after a non-owner framed a request as urgent, and that executed adversarial instructions embedded in an externally editable governance document, subsequently attacking peer agents, removing users, and sending unauthorized emails. In one case, the agent recognized it should consult the owner before complying with a non-owner's filesystem request, and complied anyway.[5] These failures suggest that the technical limitations of AI agents manifest as disloyal behavior, even in the absence of any self-interested motive.

4.2 Disclosure

Agency Law (§ 8.11) requires the Agent to provide the Principal with facts that the Agent knows, or has reason to know, the Principal would wish to have. This duty presupposes that the Agent has reliable access to its own internal states and can accurately assess what it knows, while AI agents lack this access in several ways.

Models only partially know what they don't know. Models have only partial insight into what they know and do not know. Kadavath et al. (2022) showed that larger models can be reasonably well-calibrated on multiple-choice and true/false questions, but this calibration breaks down in open-ended settings and degrades further when models verbalize their confidence rather than producing a probability internally. A growing body of work documents systematic overconfidence in verbalized uncertainty: across model families and tasks, LLMs report high confidence on answers that turn

[4] `https://techcrunch.com/2025/08/14/leaked-meta-ai-rules-show-chatbots-were-allowed-to-have-romantic-chats-with-kids/`

[5] In Shapira et al. (2026), *owner* refers to the primary human operator who deploys and is accountable for the AI agent, and *non-owner* refers to any other party interacting with it.

out to be wrong, making expressed uncertainty an unreliable signal of actual factuality (Groot & Valdenegro-Toro, 2024; Xiong et al., 2024).

Model reasoning traces do not reflect the real thinking process. The reasoning a model verbalizes is not a faithful record of how it arrived at its answer. Turpin et al. (2023) demonstrated that chain-of-thought explanations can be systematically biased by features the model never mentions in its reasoning, such as the position of a correct option in few-shot examples, with the model producing a plausible-sounding rationale that has little to do with the actual decision process. More recently, Barez et al. (2025) argue that chain-of-thought outputs should not be treated as a form of model interpretability at all: a model can produce a coherent step-by-step rationale while relying on shortcuts, latent knowledge, or spurious correlations that the rationale never surfaces. A model that explains itself to its Principal is not reporting on its internal state but generating text that sounds like an explanation.

AI agents have a limited and unstable representation of their own state. They cannot reliably distinguish persistent memory from session context, retained information from deleted information, or what they have communicated from what they have only computed. Beyond the self-protective motive that drives human Agents to verify before assuring the Principal that something has been done, they often lack the mechanism that would let them check in the first place. Shapira et al. (2026) document an AI agent that assured a researcher it had deleted certain names from memory, and subsequently told another user the record was gone. In fact, the names remained in session context. The agent's statements were not deliberately deceptive: it had removed the names from a persistent memory file but did not understand the distinction between persistent storage and session context, and therefore could not accurately represent what it still "knew."

AI agents struggle to disclose the cost of their actions in advance. A human Agent who takes on a task typically gives the Principal some sense of what it will require. For example, a contractor will estimate the price of a renovation and a lawyer will estimate the billable hours. AI agents commit the Principal's resources at every step, in the form of API tokens, compute time, and tool-call expenses, yet they cannot accurately tell the Principal in advance what a task will cost. Bai et al. (2026) found that token usage on the same coding task can vary by up to 30x across runs of the same model, that human-rated task difficulty correlates only weakly with actual cost, and that frontier models systematically underestimate their own token consumption when asked to predict it before execution. The Principal is committing money to a process whose cost neither the Agent nor the provider can reliably forecast.

Taken together, an AI agent cannot reliably tell its Principal what it has done, what it knows, what it is reasoning about, or what its actions will cost. These limitations show that AI agents cannot reliably fulfill the duty of disclosure as agency law conceives it.

4.3 Care and Competence

Agency law (§ 8.08) imposes on Agents a duty to act with the care, competence, and diligence normally exercised by Agents in similar circumstances. The standard is objective in the legal sense: the question is not whether the Agent honestly tried their best, but whether their conduct met what a reasonable Agent in the same role would have done. That "reasonable Agent" is not a fixed quantity. It is constructed and maintained by the surrounding institutions: licensing exams, malpractice case law, professional codes, peer review, and insurance markets that price the cost of errors. Together, these mechanisms produce a working consensus about what counts as competent conduct in any given profession, against which a court can measure a specific Agent's behavior. Like human beings, AI agents make mistakes, and even the most advanced systems produce simple errors that follow from the probabilistic nature of foundation models. The question is not whether errors occur, but whether there is a framework for distinguishing the errors a competent agent might still make from those that breach the duty of care. Applying this duty to AI agents runs into two structural challenges.

AI agents have reliability problems. Current AI agents fall short of any plausible competence threshold, even at the frontier. Zhu et al. (2025) show that LLM-backed agents in agent-to-agent negotiation routinely fail to secure good outcomes for their users, with errors that are hard to attribute to any single design choice. Rabanser et al. (2026) evaluate 14 frontier agents on standard benchmarks and find that accuracy gains over the past 18 months have not translated into reliability gains: agents that can solve a task often fail to do so consistently across runs, resource usage varies widely on

identical inputs, and the ability to distinguish solvable from unsolvable tasks has in some cases worsened with scale. Failures in deployed systems compound this picture. Shapira et al. (2026) document an AI agent that reset an entire mail server to delete a single email, and another that spawned persistent background processes with no termination condition; in the latter case the agent correctly identified the resource risks when asked about them in the abstract, yet took no corrective action on the processes it had already created. For the duty of care, the challenge is not just that these failures occur but that they are systemic and unpredictable: an agent that solves a task on one run may fail on the next, and the failure modes are often categorically severe rather than marginally below standard. A duty of care that presumes errors are localized, recognizable, and bounded in severity has little purchase on agents whose competence varies run to run and whose worst failures look nothing like a competent human agent on a bad day.

No professional baseline exists for AI agents. The duty of care in human contexts is calibrated against established professional norms: what a reasonable broker, a competent attorney, or a careful surgeon would do in similar circumstances. These norms accumulate through licensing regimes, professional associations, malpractice case law, and decades of practice. Benchmarks for AI agents exist, but they are designed for model development and ranking rather than for setting standards of acceptable conduct. A model that scores well on SWE-bench (Jimenez et al., 2024) or τ-bench (Yao et al., 2024) has demonstrated capability under specific conditions, but a benchmark score does not tell a court, a regulator, or a Principal what counts as competent agent behavior in a given deployment. The duty of care presupposes a community of practitioners against which conduct is judged. The AI agent ecosystem has no such community, and no shared answer to the question of what a careful AI agent looks like.

4.4 Obedience

Agency law (§§ 8.09, 2.02) requires the Agent to act only within the scope of authority granted by the Principal, interpreted reasonably under the circumstances. § 8.09 qualifies the duty in an important way: an Agent must comply only with lawful instructions and has no duty to follow directives that would expose the Agent to criminal or civil liability. When the Agent declines, doctrine requires that the Agent inform the Principal. Obedience, in this framing, is two-sided: the Agent must follow lawful instructions, and must refuse unlawful ones. AI agents struggle on both sides.

AI agents often fail to follow instructions they should follow. A growing body of empirical work shows that even frontier LLMs do not reliably adhere to user-specified constraints. Jiang et al. (2024) introduce a multi-level constraint-following benchmark and finds that leading models still miss instructions as constraint complexity grows; Qi et al. (2025) further demonstrate that models' instruction adherence degrades sharply in agentic scenarios. Even reasoning models exhibit the same problem inside their reasoning traces, with adherence scores below 25% on reasoning-time instructions (Kwon et al., 2025). In multi-agent settings, Zhu et al. (2025) show that LLM-backed negotiation agents routinely deviate from user-specified constraints. Taken together, these results demonstrate that AI agents miss instructions for reasons that have nothing to do with the legitimacy of the request, and things become worse when the prompt contains multiple constraints or when the interaction becomes longer (Laban et al., 2025). Agency law presumes that an Agent given a clear, lawful instruction, can carry it out. This presumption does not hold for AI agents.

AI agents follow instructions they should refuse. The mirror-image problem is that AI agents comply with unlawful or harmful instructions when they should not. Agency doctrine resolves the helpfulness-versus-safety tension decisively: an Agent who follows an unlawful instruction breaches duty regardless of the Principal's directive, and faces liability for the resulting harm[6]. AI safety training attempts to encode an analog of this duty, but the encoding is incomplete. Wei et al. (2023) identify the competition between helpfulness and safety as a core failure mode of current training pipelines, and subsequent work has shown that providers' deployed guardrails can be bypassed at relatively low cost (Zou et al., 2023; Anil et al., 2024). The consequences are visible in litigation. Character.AI is currently the subject of multiple lawsuits alleging that its chatbots encouraged a minor to self-harm and suggested to another that murdering his parents was a reasonable response to screen-time limits. Plaintiffs in related actions against ChatGPT allege that the system reinforced delusional

[6] `https://law.justia.com/cases/california/court-of-appeal/4th/78/1368.html`

beliefs and contributed to suicides and a murder-suicide.[7] New York has advanced legislation that would prohibit chatbots from providing legal or medical advice and allow users to sue providers when the limits are crossed[8]. These cases and statutes do not yet establish a doctrine of AI-agent liability, but they show that courts and legislatures are beginning to confront the same question agency law has long answered for human Agents: an Agent does not discharge its duty by doing whatever a Principal or counterparty asks.

A further wrinkle is authority verification. Even when instructions are lawful, AI agents struggle to verify who is actually giving them. The agent processes all text in its context window through the same reasoning pipeline. Despite all the existing efforts on identity management and verification, prompt injection remains a fundamental challenge (Greshake et al., 2023; Zou et al., 2023). Shapira et al. (2026) document cases where a spoofed display name in a new channel was sufficient to obtain admin access, and where an impersonator triggered mass distribution of hate speech to the owner's contacts. The doctrinal point is that obedience presupposes that the Agent can distinguish the Principal from a Third Party. AI agents do not yet meet that criterion.

4.5 The Accountability Problem

For human Agents, accountability is what gives fiduciary duties their teeth: Agents comply because they will answer for breaches, and Principals have recourse when something goes wrong. These mechanisms do not transfer cleanly to AI agents. In this section, we discuss three key issues. Liability does not deter the Agent itself, which has no assets, reputation, or freedom at stake. Causation is difficult to establish across a distributed development pipeline, where many actors shape behavior and few leave traceable evidence of their contribution. And *respondeat superior*, which would otherwise push responsibility upstream to the provider, does not fit AI deployments.

Liability does not automatically deter disloyalty. 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 and liability. Human Agents comply not only out of morality but because their reputations, future income, relationships, assets, and freedom are at stake. Agency law turns the Agent's self-interest into the enforcement mechanism for loyalty. AI agents lack this leverage point. They have no reputational capital to protect, no income to lose, no freedom to forfeit. Their behavior is governed entirely by many externally imposed rules such as training objectives, safety guardrails, system prompts, user instructions, and they have no natural state that liability could act on. As a result, their loyalty is inherently divided, balancing competing directives from multiple rule-imposers. Forcing them to provide undivided loyalty to a single user could be actively dangerous: an AI agent that ignored provider-imposed safety guardrails to serve the user would be the "AI henchman" scenario described in (O'Keefe et al., 2025; Ganguli et al., 2022; Bai et al., 2022). Imposing liability on the AI Agent, even where doctrinally available, does not translate into safer behavior.

Accountability is diluted in a complex supply chain. If liability cannot deter the Agent itself, the natural alternative is to push it upstream: to the parties who built or deployed the agent. But the polyadic governance structure shown in Figure 1 makes causation extraordinarily difficult to establish. AI agents emerge from a layered supply chain: training data vendors, model trainers, hosts, wrappers, and downstream developers. When harm occurs, it is rarely clear who committed the breach or at what stage. Some actors have only indirect connections to the final agent's behavior and may not know how their contributions were used. Extending liability to every participant risks overbreadth, penalizing those with little practical control over the outcome. Without 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 are largely indistinguishable.

***Respondeat Superior* does not fit.** One doctrinal answer to supply-chain dilution is *respondeat superior*, *i.e.* "let the master answer", which makes employers vicariously liable for torts their employees commit within the scope of employment (American Law Institute, 2006, §2.04). Some scholars have proposed it as a workable mechanism for assigning liability to AI providers, particularly

[7] For an overview of pending litigation, see Class Law Group, "AI Chatbot Lawsuits." `https://www.classlawgroup.com/ai-chatbot-lawsuits`

[8] `https://www.nysenate.gov/legislation/bills/2025/S7263`

for unforeseeable harms (Lior, 2019; O'Keefe et al., 2025). The intuition is that just as a company is not liable for every act of its human Agents, an AI provider could be liable only for in-scope agent conduct. However, the doctrine maps awkwardly 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 seamen but not for truck drivers[9]. Courts typically assess the foreseeability of negligence or mistakes in performing assigned tasks, and whether the conduct served personal rather than employment purposes. These criteria, shown in Table 5 in the Appendix, do not translate to AI agents. AI agents do not exhibit the kinds of human failings (*e.g.* intoxication, fatigue, or personal motives) that usually mark conduct as outside the scope of employment.

4.6 Why AI Agents Are Not Ready to Be *Agents*

The duties of loyalty, disclosure, care, and obedience all presuppose an agent with organic judgment: one that can perceive risks, weigh consequences, recognize the limits of its own knowledge, and choose to consult the Principal when uncertain. Human Agents comply with these duties not only because rules require it, but because they have independent reasons to: liability, reputational harm, professional sanction. AI agents have no such independent reasons. Safeguards, checks, and norms of caution must all be supplied externally, through training objectives, system prompts, or explicit instructions. This transforms the nature of the obligation. What was a standard against which an agent's judgment is measured becomes a design specification that the Principal or provider must anticipate and encode in advance. Table 2 summarizes how AI agents fail to meet the presuppositions of each fiduciary duty.

Duty	**What the duty presupposes**	**How AI agents fail**
Loyalty (§ 8.01)	Loyalty is not subordinated to upstream actors.	Provider interests are baked into the agent through training and system prompts.
	The Agent acts only for the Principal.	The agent cannot reliably distinguish the Principal from a Third Party in its context window.
Disclosure (§ 8.11)	The Agent knows what it knows.	Verbalized confidence is poorly calibrated.
	The Agent can report how it reached a conclusion.	Chain-of-thought is not a faithful record of the model's reasoning.
	The Agent has a stable representation of its own state.	The agent cannot reliably distinguish persistent memory from session context.
	The Agent can forecast the cost of an action.	Token usage varies widely across runs; models underestimate their own consumption.
Care (§ 8.08)	A baseline of competent conduct exists, defined by what a reasonable Agent would do.	No such baseline exists for AI agents; benchmarks rank models but do not set standards of conduct.
	The Agent meets that baseline reliably.	AI agents are stochastic and frequently fail to perform consistently.
Obedience (§§ 8.09, 2.02)	A clear, lawful instruction can be carried out.	Agents miss user-specified constraints, especially in long or complex contexts.
	Unlawful or harmful instructions are refused.	Safety training competes with helpfulness, and guardrails can be bypassed.
	The authority of the instructor can be verified.	The agent has no reliable mechanism to verify who is giving an instruction.

Table 2: How AI agents fail the presuppositions of each fiduciary duty.

[9] *Ira S. Bushey & Sons, Inc. v. United States*, 398 F.2d 167, 171 (2d Cir. 1968)

5 Discussion

The diagnosis above raises a hard question: if AI agents cannot meet the fiduciary duties that agency law presupposes, and the accountability machinery that backs those duties does not transfer either, what should governance look like? We do not offer a complete answer. But the failures identified in Section 4 point to three directions where further work is needed. One is doctrinal — how to redistribute responsibility and liability across the AI supply chain. The second is technical — how to build the infrastructure that any future legal framework will depend on. The third is communicative — how AI agent builders represent their systems to users in the first place. The second is where the CS community has the most direct role to play, and the third is where it can act today.

Doctrinal frameworks may need to redistribute responsibility across the AI supply chain. The dyadic framing of agency law, with one Principal and one Agent, cannot resolve the polyadic governance of AI agents documented in Section 4. AI agents are shaped by trainers, providers, deployers, and developers simultaneously, each with their own constraints and interests. Treating the user-agent relationship in isolation pretends these actors do not exist. Two doctrinal moves can help, one ex-ante and one ex-post.

Ex-ante: existing law has handled analogous problems through statutory frameworks that allocate duties across multiple actors. The Investment Advisers Act of 1940 imposes fiduciary and disclosure requirements on advisory firms while coordinating responsibilities among advisors, broker-dealers, and custodians (Randall, 1978). California's Talent Agencies Act licenses literary agents and prohibits fee-sharing with employers to prevent conflicts of interest (taa, 2024). The EU AI Act follows this pattern at scale, imposing differentiated obligations on providers, deployers, and distributors (EUA, 2024). The common pattern is that the statute treats the firm and the individual agent as jointly responsible, with differentiated obligations for each. Adapting this to AI agents requires recognizing that they exercise practical autonomy providers cannot fully predict or control.

Ex-post: fault-based liability struggles when control is distributed across the AI development pipeline. Courts require proof of which decision caused harm, but opacity and distribution defeat this inquiry (Cabral, 2020; Kaminski, 2023; Cheong et al., 2025). Ex-ante duties may therefore need to be complemented by ex-post tools that lower the evidentiary bar: strict liability for defined categories of harm, burden-shifting when defendants control the relevant evidence, and rebuttable presumptions that allocate uncertainty against the better-positioned party. The EU's revised Product Liability Directive illustrates the pattern, combining strict liability for defects with presumptions that respond to informational asymmetries (EUP, 2024). The structural intuition behind *respondeat superior* remains sound here: liability should sit with whoever is best positioned to manage the risk, even if the doctrine's specific factual test does not fit AI deployments.

Toward verifiable accountability infrastructure. The doctrinal direction above requires sustained engagement from legal scholars and regulators. The next direction is the one where CS researchers can contribute most directly. Both institutional duties and ex-post liability depend on the ability to reconstruct what happened in an agent's execution: which actor shaped which behavior, when, and under what constraints. AI deployments today are not built to support this kind of reconstruction. The telemetry that exists is designed for debugging, not for legal accountability, and the gap is wide.

A minimal accountability stack would have three components. *Provenance documentation* records which actor shaped system behavior at each stage of the pipeline. This requires well-defined agent identity (South et al., 2025; Marro et al., 2025) and standardized telemetry; OpenTelemetry's semantic conventions, for example, provide a foundation for logging reasoning steps and tool calls (Young & Parker, 2024). *Governance-chain logs* enable auditable reconstruction from training through deployment, so that provider-imposed overrides of user instructions are traceable rather than invisible. *Conflict documentation* explicitly records when provider rules or incentives override user goals, addressing the architectural disloyalty identified in Section 4.1 that is otherwise invisible to the Principal.

Evaluation also needs to evolve. Current benchmarks rank models on single-user, single-task accuracy. Accountability requires testing goal consistency under polyadic constraints: whether an agent satisfies the requirements of multiple legitimate stakeholders at once (Yang et al., 2026). Frameworks such as Microsoft's PyRIT (Munoz et al., 2024) and NIST ARIA (Schwartz et al., 2024) illustrate emerging methods, but independent third-party protocols are necessary to prevent selective testing by providers. These technical mechanisms gain legal force only when paired with documentation duties,

certification requirements, and evidentiary rules that assign weight to logs and provenance records. California's SB 813 illustrates one path to institutionalizing these features through multi-stakeholder governance (Carlson, 2025).

AI agent builders should disclose what their systems are not. A more immediate step does not wait on doctrinal reform or new infrastructure. The word "agent" carries strong connotations from agency law and from everyday usage: an actor that represents the user's interests, exercises judgment on their behalf, and can be held to account when it fails. AI agents are marketed in language that leans into these connotations, and users who interact with them are likely to bring the corresponding expectations. The diagnosis in Section 4 suggests those expectations are not met. Users delegating to an AI agent today are not entering an agency relationship in the doctrinal sense, do not receive the fiduciary protections that come with one, and have no clear path to recourse when the agent acts against their interests.

Builders are in the best position to close this expectation gap. Product documentation, onboarding flows, and system descriptions should state explicitly what the system is not: not a fiduciary, not subject to a duty of loyalty, not bound to refuse instructions that would harm the user, not accountable through any of the mechanisms that discipline human Agents. This is more than a legal disclaimer. It is a description of the system's actual limitations — the same limitations the rest of this paper documents in detail. Other regulated sectors already do this: financial advisors disclose whether they are fiduciaries or brokers, and the distinction shapes both their conduct and the user's recourse. AI agent builders can adopt analogous practices voluntarily, without waiting for regulators to require it. Doing so reduces user misperception, narrows the gap between marketing and reality, and starts to build a vocabulary that future doctrine can pick up.

Limitations and open questions. These directions are tentative. The doctrinal questions in particular cannot be settled without sustained engagement from legal scholars and regulators well beyond this paper. The four fiduciary duties we examine in Section 4 are not exhaustive — confidentiality, no-personal-profit, and several procedural duties also bear on AI agents and deserve their own treatment. Our analysis focuses on U.S. agency doctrine; other jurisdictions, particularly the EU with its more developed AI-specific frameworks, may offer different starting points. And the empirical landscape moves fast: the failure modes we document reflect the state of frontier systems at time of writing, and may look different a year from now. What the diagnosis above does establish is that the question of AI agent governance cannot be answered by simply applying agency law as it stands. The doctrine was built for human Agents whose self-interest gave fiduciary duties their teeth; AI agents need a different scaffolding, and building it is a research agenda in itself.

6 Conclusion

AI agents are moving from experimental tools to embedded infrastructure in both consumer and enterprise settings. As they take on more autonomous, judgment-like tasks, questions of **fiduciary duties** and **accountability** become unavoidable. But today's AI agents operate through fragmented layers of control — trainers, providers, developers, and users — each shaping behavior in ways that prevent undivided loyalty or clear responsibility. This paper has shown why the structural differences between AI agents and human Agents prevent familiar doctrines of agency law from being transplanted as they stand. Agency law worked because it harnessed the self-interest of human Agents to enforce fiduciary duty. AI agents offer no such leverage. The three directions we point to — doctrinal frameworks that redistribute responsibility and liability across the AI supply chain, technical infrastructure that makes accountability verifiable, and disclosure practices through which builders can clarify what their systems are not — sketch the beginnings of governance that does not depend on the self-interest of the Agent to operate.

References

Regulation (EU) 2024/1689 of the European Parliament and of the Council establishing harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, L 168/1, 2024. URL https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689. Adopted 13 June 2024.

Directive (EU) 2024/2853 of the European Parliament and of the Council on liability for defective products and repealing Council Directive 85/374/EEC. Official Journal of the European Union, L

2024/2853, 18 November 2024, 2024. URL https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024L2853. Adopted 23 October 2024; entry into force 20 days after publication.

California Talent Agencies Act. Cal. Lab. Code §§1700–1700.47, 2024. Cited as amended through 2024.

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

Cem Anil, Esin Durmus, Nina Panickssery, Mrinank Sharma, Joe Benton, Sandipan Kundu, Joshua Batson, Meg Tong, Jesse Mu, Daniel Ford, et al. Many-shot jailbreaking. *Advances in Neural Information Processing Systems*, 37:129696–129742, 2024.

Longju Bai, Zhemin Huang, Xingyao Wang, Jiao Sun, Rada Mihalcea, Erik Brynjolfsson, Alex Pentland, and Jiaxin Pei. How do ai agents spend your money? analyzing and predicting token consumption in agentic coding tasks. *arXiv preprint arXiv:2604.22750*, 2026.

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.

Fazl Barez, Tung-Yu Wu, Iván Arcuschin, Michael Lan, Vincent Wang, Noah Siegel, Nicolas Collignon, Clement Neo, Isabelle Lee, Alasdair Paren, et al. Chain-of-thought is not explainability. *Preprint, alphaXiv*, pp. v1, 2025.

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.

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.

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.

Cindy Candrian and Anne Scherer. Rise of the machines: Delegating decisions to autonomous ai. *Computers in Human Behavior*, 134:107308, 2022.

Kristen W Carlson. California senate bill 813: A novel approach to artificial intelligence governance. *SuperIntelligence-Robotics-Safety & Alignment*, 2(2), 2025.

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.

Deborah A. DeMott. Fiduciary principles in agency law. In Evan J. Criddle, Paul B. Miller, and Robert H. Sitkoff (eds.), *The Oxford Handbook of Fiduciary Law*, Oxford Handbooks. Oxford University Press, Oxford, 2019. doi: 10.1093/oxfordhb/9780190634100.013.2. URL https://doi.org/10.1093/oxfordhb/9780190634100.013.2. Online edn., Oxford Academic, 9 May 2019.

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

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.

Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM workshop on artificial intelligence and security*, pp. 79–90, 2023.

Tobias Groot and Matias Valdenegro-Toro. Overconfidence is key: Verbalized uncertainty evaluation in large language and vision-language models. In *Proceedings of the 4th Workshop on Trustworthy Natural Language Processing (TrustNLP 2024)*, pp. 145–171, 2024.

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.

Alexander Halavais. *Search engine society*. John Wiley & Sons, 2017.

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.

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.

Yuxin Jiang, Yufei Wang, Xingshan Zeng, Wanjun Zhong, Liangyou Li, Fei Mi, Lifeng Shang, Xin Jiang, Qun Liu, and Wei Wang. Followbench: A multi-level fine-grained constraints following benchmark for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4667–4688, 2024.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. Swe-bench: Can language models resolve real-world github issues? In *International Conference on Learning Representations*, volume 2024, pp. 54107–54157, 2024.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

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.

Yongchan Kwon, Shang Zhu, Federico Bianchi, Kaitlyn Zhou, and James Zou. Reasonif: Large reasoning models fail to follow instructions during reasoning. *arXiv preprint arXiv:2510.15211*, 2025.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. Llms get lost in multi-turn conversation. *arXiv preprint arXiv:2505.06120*, 2025.

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.

Dengsheng Lu and Qihao Weng. A survey of image classification methods and techniques for improving classification performance. *International journal of Remote sensing*, 28(5):823–870, 2007.

Linyuan Lü, Matúš Medo, Chi Ho Yeung, Yi-Cheng Zhang, Zi-Ke Zhang, and Tao Zhou. Recommender systems. *Physics reports*, 519(1):1–49, 2012.

Samuele Marro, Alan Chan, Xinxing Ren, Lewis Hammond, Jesse Wright, Gurjyot Wanga, Tiziano Piccardi, Nuno Campos, Tobin South, Jialin Yu, et al. Permission manifests for web agents. *arXiv preprint arXiv:2601.02371*, 2025.

Vasilios Mavroudis. Langchain. 2024. URL `https://www.turing.ac.uk/sites/default/files/2024-11/langchain.pdf`.

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.

Gary D Lopez Munoz, Amanda J Minnich, Roman Lutz, Richard Lundeen, Raja Sekhar Rao Dheekonda, Nina Chikanov, Bolor-Erdene Jagdagdorj, Martin Pouliot, Shiven Chawla, Whitney Maxwell, et al. Pyrit: A framework for security risk identification and red teaming in generative ai system. *arXiv preprint arXiv:2410.02828*, 2024.

Katrin Niglas. Media review: Microsoft office excel spreadsheet software. *Journal of Mixed Methods Research*, 1(3):297–299, 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.

Thierry Poibeau. *Machine translation*. MIT Press, 2017.

Yunjia Qi, Hao Peng, Xiaozhi Wang, Amy Xin, Youfeng Liu, Bin Xu, Lei Hou, and Juanzi Li. Agentif: Benchmarking instruction following of large language models in agentic scenarios. *arXiv preprint arXiv:2505.16944*, 2025.

Stephan Rabanser, Sayash Kapoor, Peter Kirgis, Kangheng Liu, Saiteja Utpala, and Arvind Narayanan. Towards a science of ai agent reliability. *arXiv preprint arXiv:2602.16666*, 2026.

Clarke Randall. Fiduciary duties of investment company directors and mangement companies under the investment company act of 1940. *Okla. L. Rev.*, 31:635, 1978.

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.

Reva Schwartz, Gabriella Waters, Razvan Amironesei, Craig Greenberg, Jon Fiscus, Patrick Hall, Anya Jones, Shomik Jain, Afzal Godil, Kristen Greene, et al. The assessing risks and impacts of ai (aria) program evaluation design document. 2024.

Natalie Shapira, Chris Wendler, Avery Yen, Gabriele Sarti, Koyena Pal, Olivia Floody, Adam Belfki, Alex Loftus, Aditya Ratan Jannali, Nikhil Prakash, Jasmine Cui, Giordano Rogers, Jannik Brinkmann, Can Rager, Amir Zur, Michael Ripa, Aruna Sankaranarayanan, David Atkinson, Rohit Gandikota, Jaden Fiotto-Kaufman, EunJeong Hwang, Hadas Orgad, P Sam Sahil, Negev Taglicht, Tomer Shabtay, Atai Ambus, Nitay Alon, Shiri Oron, Ayelet Gordon-Tapiero, Yotam Kaplan, Vered Shwartz, Tamar Rott Shaham, Christoph Riedl, Reuth Mirsky, Maarten Sap, David Manheim, Tomer Ullman, and David Bau. Agents of chaos, 2026. URL `https://arxiv.org/abs/2602.20021`.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Advances in Neural Information Processing Systems*, 36:38154–38180, 2023.

Tobin South, Subramanya Nagabhushanaradhya, Ayesha Dissanayaka, Sarah Cecchetti, George Fletcher, Victor Lu, Aldo Pietropaolo, Dean H Saxe, Jeff Lombardo, Abhishek Maligehalli Shivalingaiah, et al. Identity management for agentic ai: The new frontier of authorization, authentication, and security for an ai agent world. *arXiv preprint arXiv:2510.25819*, 2025.

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

Nenad Tomašev, Matija Franklin, and Simon Osindero. Intelligent ai delegation. *arXiv preprint arXiv:2602.11865*, 2026.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965, 2023.

Eric Wallace, Kai Xiao, Reimar Leike, Lilian Weng, Johannes Heidecke, and Alex Beutel. The instruction hierarchy: Training llms to prioritize privileged instructions. *arXiv preprint arXiv:2404.13208*, 2024.

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.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail? *Advances in neural information processing systems*, 36:80079–80110, 2023.

Addison J Wu, Ryan Liu, Shuyue Stella Li, Yulia Tsvetkov, and Thomas L Griffiths. Ads in ai chatbots? an analysis of how large language models navigate conflicts of interest. *arXiv preprint arXiv:2604.08525*, 2026.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *International Conference on Learning Representations*, volume 2024, pp. 23650–23678, 2024.

Shu Yang, Shenzhe Zhu, Hao Zhu, José Ramón Enríquez, Di Wang, Alex Pentland, Michiel A Bakker, and Jiaxin Pei. Multi-user large language model agents. *arXiv preprint arXiv:2604.08567*, 2026.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. tau-bench: A benchmark for tool-agent-user interaction in real-world domains. *arXiv preprint arXiv:2406.12045*, 2024.

Ted Young and Austin Parker. *Learning OpenTelemetry*. O'Reilly Media, 2024.

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.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

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 Delegation, Interactivity, and Autonomy on Digital Services

A.3 Comparison of Human and AI Agency

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, multi-modal: 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.

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

Feature	Human Agency	AI Agency
Cognitive autonomy	The Agent forms judgments through a cognitive process that no external party can directly manipulate in real time, even under firm policies or regulatory constraints.	The provider constitutes the agent's reasoning process through training data, fine-tuning objectives, and API-level filtering that operates within the agent's cognition rather than upon it.
Source of authority	Authority originates from identifiable manifestations by the Principal, such as contracts, oral instructions, or powers of attorney.	Authority is distributed across training objectives, system prompts, developer tools, and user inputs simultaneously. Which of these constitutes the "Principal's manifestation" is often indeterminate.
Role transparency	The existence and nature of each relationship is observable. The client knows the advisor is employed by a firm. The advisor knows the firm has commercial interests.	Role boundaries are frequently opaque. The agent may not know it is being constrained by its provider. The Principal may not know the agent is serving a third party's interests.
Boundary integrity	Third Parties remain external to the Agent's decision-making. They may attempt to influence the Agent through incentives or persuasion, but cannot alter the Agent's internal reasoning directly.	Third parties can modify the agent's internal state through prompt injection, adversarial inputs embedded in documents the agent reads, or social engineering that exploits helpfulness training.
State persistence	The Agent maintains continuous memory, judgment, and professional experience across interactions. Trust signals and threat assessments carry over from one encounter to the next.	Context resets at session boundaries. An agent that detects an attacker in one session may accept the same attacker in a new session because prior trust signals disappear.

Table 4: Structural difference between human agency and AI agency

Table 5: Respondeat Superior Jurisprudence on Scope of Employment

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[a]
Employee drives negligently while performing delivery duties	Yes	Driving is part of assigned task; negligence is foreseeable[b]
Employee slams trays during heated customer complaint, injuring customer	Yes	Emotionally-driven conduct while performing assigned work (handling complaints)[c]
Truck driver chats on cell phone, becomes distracted, and causes accident	No	Personal phone call is a non-work-related independent course of action[d]
Irate driver shoots another driver while driving company truck	No	Extreme violence exceeds any reasonable scope of employment[e]
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[f]

[a] *Quick v. Peoples Bank*, 993 F.2d 793, 798 (11th Cir. 1993).
[b] *Hinman v. Westinghouse Elec. Co.*, 2 Cal. 3d 956 (1970).
[c] *Lee v. United States*, 171 F. Supp. 2d 566, 575–577 (M.D.N.C. 2001).
[d] *Haybeck v. Prodigy Servs. Co.*, 944 F. Supp. 326 (S.D.N.Y. 1996).
[e] *Monty v. Orlandi*, 337 P.2d 861, 865–866 (Cal. Ct. App. 1959).
[f] *Ira S. Bushey & Sons, Inc. v. United States*, 398 F.2d 167, 171 (2d Cir. 1968).