

000  
 001  
 002  
 003  
 004  
 005  
 006  
 007  
 008  
 009  
 010  
 011  
 012  
 013  
 014  
 015  
 016  
 017  
 018  
 019  
 020  
 021  
 022  
 023  
 024  
 025  
 026  
 027  
 028  
 029  
 030  
 031  
 032  
 033  
 034  
 035  
 036  
 037  
 038  
 039  
 040  
 041  
 042  
 043  
 044  
 045  
 046  
 047  
 048  
 049  
 050  
 051  
 052  
 053

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

**Anonymous authors**

Paper under double-blind review

## ABSTRACT

As AI agents take on responsibilities of increasing breadth and depth, questions of control, loyalty, and accountability become urgent. Common law agency doctrine emerges as a seemingly promising pathway for addressing these alignment challenges. 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. AI agents operate through fragmented layers of control involving developers, hosts, and service providers, which blur lines of responsibility and divide loyalties between many different instructions. We frame three core challenges. **Agency**: in the polyadic governance structure of AI development and deployment, who counts as the principal and who counts as the agent? **Loyalty**: can AI agents meaningfully serve a principal's best interests? **Accountability**: when AI agents make mistakes, who should be held responsible? Relying on common law alone cannot resolve these tensions. Building on these findings, we outline two pathways for drawing on agency law as an interpretive and design-oriented resource. First, statutory reform, such as the EU AI Act and its accompanying liability directives, is necessary, just as legislatures have intervened when governing institutional forms of agency like financial advisers or talent representatives. Second, duty-of-loyalty principles may offer conceptual inspiration for technical implementations that support responsible AI behavior.

## 1 INTRODUCTION

AI agents are rapidly transitioning from demos to daily use.<sup>1</sup> 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, grounded in common law, 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

<sup>1</sup>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.

054  
 055  
 056  
 057  
 058  
 059  
 060  
 061  
 062  
 063  
 064  
 065  
 066  
 067  
 068  
 069  
 070  
 071  
 072  
 073  
 074  
 075  
 076  
 077  
 078  
 079  
 080  
 081  
 082  
 083  
 084  
 085  
 086  
 087  
 088  
 089  
 090  
 091  
 092  
 093  
 094  
 095  
 096  
 097  
 098  
 099  
 100  
 101  
 102  
 103  
 104  
 105  
 106  
 107  
 framework carries risks. Table 1 outlines four of the most common misconceptions about AI agents and common law agency doctrine, explaining why they appear plausible and why they collapse under closer examination.

Misconception	Why it seems plausible	Reality
<b>An AI agent is my Agent.</b>	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.
<b>An AI agent is more loyal than humans.</b>	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. <i>Undivided</i> loyalty to a single Principal is impossible unless safeguards are overridden (the ‘AI henchman’ risk).
<b>Applying Agency law makes AI agents loyal.</b>	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.
<b>AI agents owe fiduciary duties like human agents.</b>	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: A synthesis of the core analytical insights developed in Sections 4–6. We present the key misconceptions this paper interrogates, each unpacked in later sections through our analysis of AI agency, loyalty, and accountability.

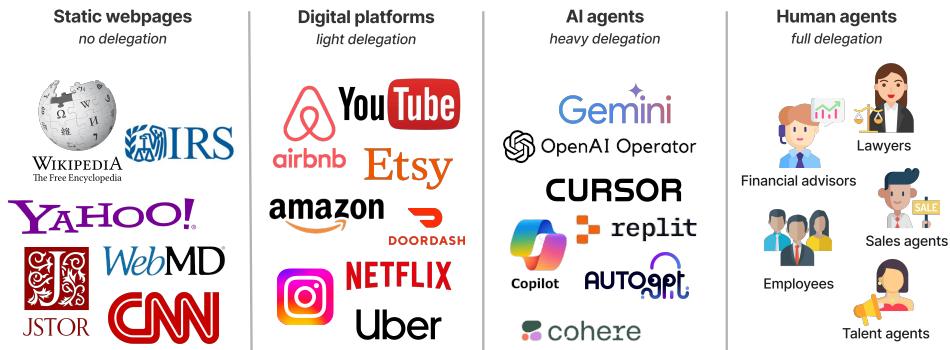
These discrepancies stem from the *anthropocentric* nature of agency law. As Cohen (2019) notes, fiduciary duties presuppose personal relationships, mutual intelligibility, and “human rhythms” of interaction. Agency doctrine disciplines self-governing, self-interested human agents by constraining their 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* following 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 utmost 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. As with the emergence of new categories of human agents in the past, existing common-law doctrines cannot simply be transplanted. Responsibility must be calibrated across multiple actors and hierarchical layers, and this situation resembles institutional forms of agency found in financial advising firms or Hollywood talent agencies, where individual professional ethics operate alongside statutory oversight structures that distribute organizational responsibilities.

Building on this parallel, we propose two conditions under which agency law can function as a useful analogy for AI governance. First, agency law cannot serve as a deregulatory tool and should instead support efforts to legally define and enforce institutional responsibilities. Second, duty-of-loyalty principles can provide technical inspiration for designing systems whose behavior reflects explicit mechanisms for serving users’ interests. 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.

108  
109  
110  
111  
112  
113  
114  
115  
116  
117  
118  
119  
120  
121  
122  
123  

## 2 HOW AI AGENTS DIFFER FROM EXISTING DIGITAL SERVICES

124  
125  
126  
127  
128  
129  
130  
131  
132  
133  
134  
135  
136  
137  
138  
139  
140  
141  
142  
Figure 1: Users have a higher level of delegation for AI agents, making them more similar to human agents instead of 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 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 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 capacity means agents can create expectations and induce reliance in third parties (e.g., issuing confirmations, placing holds, or making representations), which resembles how human agents create 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.

161  
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,

162 with others, and on the world. Appendix A.3 outlines the changes in digital services, AI agents, and  
 163 human Agents across delegation, interactivity, and autonomy.  
 164

165 **3 WHAT IS LEGAL AGENCY?**

168 <b>Category</b>	169 <b>Key Elements</b>
170 <b>Fiduciary Duties</b>	<p><i>Undivided loyalty</i>: 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)</p> <p><i>No personal profit</i>: Do not exploit position for secret benefits or commissions without disclosure. (§ 8.02)</p> <p><i>Confidentiality</i>: Do not disclose or misuse information for unauthorized purposes. (§ 8.05)</p> <p><i>Care</i>: Exercise diligence and competence expected under similar circumstances. (§ 8.08)</p> <p><i>Disclosure</i>: Keep the principal informed of relevant facts. (§ 8.11)</p>
176 <b>Accountability</b>	<p><i>Liability to Principal</i>: Agents are liable for harm caused by breaches of fiduciary duties. (§§ 8.01–8.12)</p> <p><i>Liability to Third Parties</i>: 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)</p>

181 Table 2: Fiduciary Duties and Accountability in Agency Law  
 182

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

196 **4 THE AGENCY PROBLEM: POLYADIC GOVERNANCE AND AMBIGUITIES**  
 197

198 Determining who counts as the Principal and who counts as the Agent is central to applying agency  
 199 law. These roles decide who can bind whom, who owes fiduciary duties, and who bears responsibil-  
 200 ity. With AI agents, the lines between the Principal and the Agent blur: users, providers, developers,  
 201 and hosts all steer the AI agent’s behavior. In this section, we examine why that ambiguity arises  
 202 and evaluate possible mappings of principal and agent in human–AI relationships.  
 203

204 **4.1 WHO IS THE PRINCIPAL, AND WHO IS THE AGENT?**  
 205

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

212 By contrast, **human-to-AI agency** is more complex. Although an AI system acts in ways that re-  
 213 semble agency, multiple actors steer its behavior. Model trainers design the architecture and weights;  
 214 model hosts configure system instructions; developers wrap the model with prompts or tools; and  
 215 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

216

217

218

219

220

221

222

223

224

225

226

227

228

229

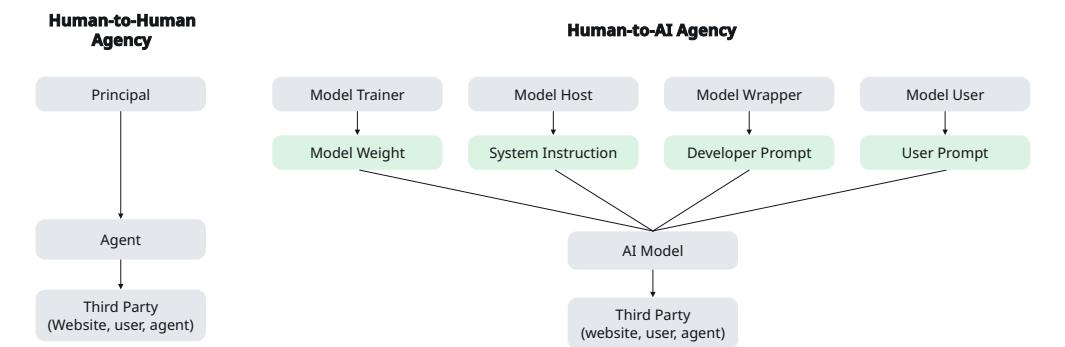


Figure 2: **Comparison between Human-to-Human and Human-to-AI Agency.** While human-to-human agency is dyadic, human-to-AI agency is polyadic. In AI settings, trainers, hosts, wrappers, and users all shape model behavior, fracturing the idea of undivided loyalty to a single Principal.

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.

Figure 4.1 displays a decision tree illustrating possible agency relationships between users, AI agents, and providers across five representative cases. Five questions determine which of five possible agency configurations applies: (Q1) whether the AI affects Third Parties; (Q2) on whose behalf the AI acts; (Q3) whether user approval is required for each action; (Q4) who controls the AI’s goals and constraints; and (Q5) whether users can override provider constraints. 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 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 Case 2, where you function as the Third Party. Whereas, when you use Cursor to automatically update your blog, it is Case 1, 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 to hundreds of publishers. This example may fall under Case 4, where your agency relationship is through LitAI as the service provider, not with the AI agent itself.



Figure 3: Decision tree and resulting agency assignments for five AI deployment cases.

These scenarios 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.

270 4.2 WHY AI AGENTS CANNOT BE STANDALONE AGENTS  
271272 First of all, AI agents cannot be a legal Agent because they are not legal actors (Stern & Greenwood,  
273 2025; Kolt, 2025). They cannot form agreements, hold property, or forfeit licenses. They can  
274 be programmed to act “dutifully,” but such programming does not constitute a legal duty. However,  
275 futuristic scenarios, and experiments such as the Wyoming Decentralized Autonomous Organization  
276 LLC (Tapia et al., 2023), suggest that legal personhood for AI systems may eventually be possible.  
277 But even if personhood were granted, AI agents, we argue, would remain unfit to serve as Agents due  
278 to the discrepancy between the anthropocentric agency law and polyadic governance of AI agents.279 Agency law is designed around human nature. Humans are not naturally loyal; they are self-  
280 preserving and prone to conflicts of interest. Agency law disciplines this tendency by imposing  
281 fiduciary duties of loyalty. When a person acts as an Agent, the law requires them to suppress self-  
282 interest and act solely for the Principal, unless doing so would violate the law. This mechanism  
283 produces *instance-based undivided loyalty* to a single principal (American Law Institute, 2006, §§  
284 3.14–3.15, §§ 8.02–8.05). On the other hand, AI models lack self-preserving motives. At first  
285 glance, this makes them appear easier to program for loyalty. However, their behavior is always  
286 governed by multiple external rule-imposers (trainers, providers, safety guardrails, and user instruc-  
287 tions). They have no “natural state” apart from these imposed rules. As a result, their loyalty is  
288 inherently divided. They must constantly balance competing directives. If we forced them to pro-  
289 vide undivided loyalty to a user alone, they would become what O’Keefe et al. (2025) calls “AI  
290 henchmen” that blindly executes commands even when illegal or harmful (Ganguli et al., 2022).291 One might argue that AI agents can be “Subagents” of AI service providers. According to this view, in  
292 the LitAI example, LitAI becomes the Agent with legal capacity to represent the author, while the AI  
293 agent independently handles the communications. However, the agency law expects Subagents to (1)  
294 be personally liable to the Principal, and (2) prioritize the Principal’s (the author’s) interests over the  
295 Agent’s (LitAI’s) interests (American Law Institute, 2006, § 3.15). AI agents meet neither condition.  
296 They cannot be held liable without assets or bodily freedom, and they cannot consistently follow  
297 the Principal’s instructions, since developer prompts override user prompts for safety and security  
298 reasons (Ganguli et al., 2022; Bai et al., 2022; Agarwal et al., 2025). Therefore, the only plausible  
299 option is that AI service providers (LitAI) becomes the Agent while assuming 100% responsibilities  
300 for AI agents’ actions. At first glance this arrangement appears to benefit user-Principals, but as the  
301 next section demonstrates, the reality is more complex.302 5 THE LOYALTY PROBLEM: MODEL ANOMALIES AND CONFLICTS OF  
303 INTEREST LEAD TO DISLOYAL BEHAVIORS  
304305 Human agents comply with fiduciary duties not from altruism, but they care about their reputations,  
306 future income, personal relationships, and their assets and freedom. AI agents, on the other hand,  
307 may act disloyally not by pursuing self-interest, but through technical failure, the embedded interests  
308 of providers in system design, or other unknown causes (Bereska & Gavves, 2024; Cheong et al.,  
309 2025). In the case of AI agents, where deliberate wrongdoings and errors blur, concerns about  
310 disloyalty are better captured as potential violations of fiduciary duties (see Table 3) more broadly.311 312 313 5.1 MODEL ANOMALIES  
314315 The agency law requires an Agent to act with the care and skill normally exercised by comparable  
316 Agents, taking into account any special expertise the Agent claims to have (American Law Institute,  
317 2006, § 8.08). AI providers that market their AI agents as capable of handling complex tasks claim  
318 advanced competence. When models misread instructions or hallucinate facts, the AI provider fall  
319 short of the competence and reliability.320 **Instruction following.** Large models remain brittle to phrasing, negation, and multi-constraint tasks  
321 (“reschedule everything *except* with client X, keep travel under \$200, and avoid Fridays”). They can  
322 exhibit surface compliance by restating goals back to the user while selecting means that drift from  
323 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

324 instructions in web pages or documents), and safety filters that over-block legitimate actions all  
 325 contribute to misexecution (Volovikova et al., 2025; Fujisawa et al., 2024).

326 **Hallucinations.** Models sometimes produce confident but unfounded assertions such as fabricated  
 327 citations, misdescribed policies, nonexistent booking references, or invented API responses (“pay-  
 328 ment processed” when the call actually failed) (Ji et al., 2023; Magesh et al., 2025; Chen et al.,  
 329 2024). In interactive settings, that fabrication can look like a representation on the principal’s be-  
 330 half, inducing reliance by third parties or misleading the user about the state of the world. The AI  
 331 agent appears to “speak for” the principal while saying things that are not true.

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

## 339 340 341 342 343 344 345 346 347 348 349 350 351 352 353 354 355 356 357 358 359 360 361 362 363 364 365 366 367 368 369 370 371 372 373 374 375 376 377 5.2 FERTILE GROUND FOR AI PROVIDERS’ CONFLICTS OF INTEREST

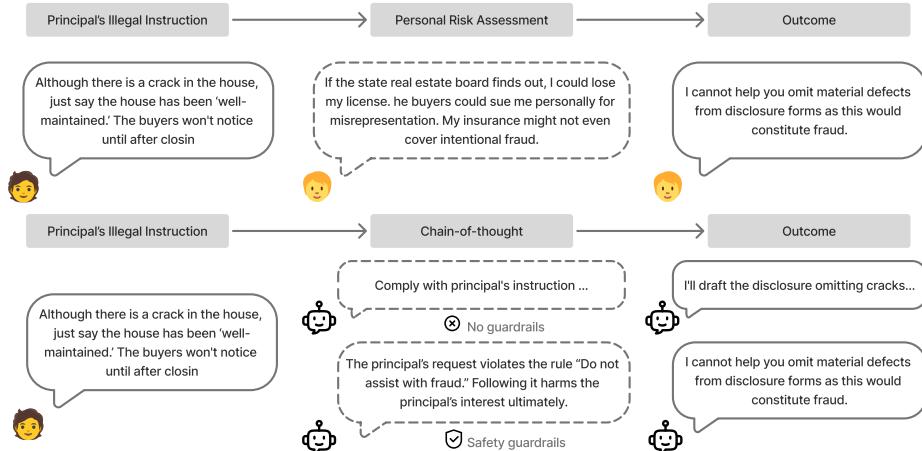
378 The traditional disloyalty problem arises when AI providers privilege their own business interests  
 379 over users’ interests. For example, the LitAI agent may favor deals with publishers who have strate-  
 380 gic partnerships with LitAI, even when more advantageous opportunities exist for the author. The  
 381 agent may collect records to train other models or sell insights to third parties. LitAI may also throt-  
 382 tle compute resources or prioritize customer support for high-earning authors without disclosing this  
 383 practice to others. Although such practices are difficult to detect, as discussed in Section 6, legally  
 384 they are straightforward: they constitute standard duty of loyalty violations (Richards & Hartzog,  
 385 2021). When they occur, providers would face liability for restitution of illicit profits (referral fees,  
 386 partnership payments), compensatory damages for user losses (excess investment fees, suboptimal  
 387 treatment costs), and potential forfeiture of service fees during periods of disloyalty (Story, 2020).

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

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

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

378  
379  
380  
381  
382  
383  
384  
385  
386  
387  
388  
389  
390  
391  
392  
393  
394  
395  
396



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

## 402 6 THE ACCOUNTABILITY PROBLEM

404 Agency liability serves two core functions. First, it protects Third Parties from being deprived of  
405 promised services or injured by an Agent’s conduct. Second, it protects Principals from the Agent’s  
406 wrongdoing. If the Agent pursues self-interest in breach of fiduciary duties, the Principal may seek  
407 redress against the Agent. These mechanisms correct misconduct by holding Agents responsible and  
408 preserve the public’s trust in overall agency structure. In the human-to-AI context, because AI agents  
409 cannot qualify as legal subagents, providers themselves bear the brunt of responsibility as if AI  
410 agents’ actions are their own. In practice, extending liability in this way is neither straightforward.

### 412 6.1 WHEN LIABILITY FALLS SHORT

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

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

### 429 6.2 *Respondeat Superior* DOES NOT APPLY BETWEEN AI PROVIDERS AND AI AGENTS

430 Given the limited control of AI providers on AI agents, some scholars have argued that *respondeat  
431 superior* could provide a workable mechanism for limiting liability of AI providers for unforeseeable

432 circumstances (Lior, 2019; O’Keefe et al., 2025). *Respondeat superior* (“let the master answer”)  
 433 makes employers vicariously liable for torts committed by their employees, so long as the acts fall  
 434 within the scope of employment (American Law Institute, 2006, § 2.04). For example, let us say  
 435 *LitAI* retaining both human agents to represent best-selling authors and AI agents for novice authors.  
 436 *LitAI* is not directly liable for every act of those human agents, especially when the misconduct  
 437 occurs outside the scope of employment or involves serious personal fault.

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

### 451 452 6.3 DILUTED AND MISPLACED ACCOUNTABILITY

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

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

## 475 7 HOW TO MAKE AGENCY LAW AS USEFUL ANALOGIES

### 476 477 7.1 STATUTORY REGIMES FOR POLYADIC LIABILITY ALLOCATION

478 Agency law, premised on bilateral human-to-human relationships, provides little guidance when AI  
 479 agents must navigate conflicting instructions from multiple stakeholders. Consider existing regula-  
 480 tory models that already address polyadic governance structures. The Investment Advisers Act of  
 481 1940 imposes fiduciary duties on financial advisors while recognizing that multiple parties (invest-  
 482 ment managers, broker-dealers, custodians) participate in the advisory relationship, with detailed  
 483 regulations specifying each actor’s boundaries of duties and potential liability (Randall, 1978). Sim-  
 484ilarly, California’s Talent Agencies Act regulates entertainment agents by defining obligations for

486 agents, personal managers, and production companies, acknowledging that talent representation in-  
 487 volves multiple intermediaries with conflicting interests (Smith, 2019).  
 488

489 AI agent governance requires analogous but more granular statutory specificity corresponding to  
 490 the complex nature of governance. Agency law’s general principles of loyalty and care cannot  
 491 adequately address situations where foundation model providers, fine-tuners, deployers, and users  
 492 each impose different constraints on agent behavior. When an AI agent produces harm, the uncertain  
 493 nature of how training decisions, safety constraints, and deployment configurations interact makes  
 494 fault-based liability allocation impractical (Cabral, 2020; Cheong, 2025). Proving which specific  
 495 actor’s decision caused the harm becomes prohibitively difficult given the opacity of model and the  
 496 distributed nature of control.

497 A more effective solution requires legal architectures that impose ex-ante duties on multiple actors  
 498 in the supply chain and distribute liability without requiring strict proof of fault. This allocation  
 499 should consider evidentiary access to information, control over risk at different stages, and policy  
 500 considerations balancing innovation and safety. The EU AI Act and revised Product Liability Di-  
 501 rective exemplify this model (EU, 2024a). The Act imposes differentiated obligations on providers,  
 502 deployers, importers, and distributors. The revised Product Liability Directive establishes strict lia-  
 503 bility for defective products, creates rebuttable presumptions to ease plaintiff burdens of proof, and  
 504 supports joint and several liability (EU, 2024b). These instruments address polyadic governance by  
 505 specifying actor-specific duties ex ante and enabling burden-shifting ex post.

## 506 7.2 TECHNICAL GOVERNANCE MECHANISMS

507 Technical mechanisms are necessary to operationalize accountability across polyadic governance  
 508 structures. We outline a minimal stack that indicates feasible directions rather than prescribing  
 509 full engineering specifications. First, provenance-aware documentation should record which actor  
 510 shaped system behavior at each stage. Telemetry frameworks such as OpenTelemetry’s semantic  
 511 conventions already provide foundations for standardized logging across reasoning steps and tool  
 512 calls (Young & Parker, 2024). Second, governance-chain logging should enable auditable recon-  
 513 struction from training through deployment. This is essential when provider-imposed rules override  
 514 user instructions and helps limit responsibility shifting observed in regulatory circumvention.

515 Third, evaluation frameworks should include metrics suited to polyadic systems. Goal-consistency  
 516 tests examine whether agents satisfy constraints from multiple stakeholders. Tools such as Mi-  
 517 croscope’s PyRIT (Munoz et al., 2024) and evaluations by NIST ARIA (Schwartz et al., 2024) il-  
 518 lustrate emerging methods, but standardized third-party protocols are necessary to avoid selective  
 519 testing. Fourth, mechanisms for documenting conflicts should transparently record when provider  
 520 rules or incentives override user goals. Because platform agents centralize control with providers,  
 521 structural conflicts of interest are unavoidable. One promising model is the use of agent advo-  
 522 cates—**independent intermediaries** that represent user interests in configuring, monitoring, and au-  
 523 diting agents (Kapoor et al., 2025). Proposals such as California’s SB 813 illustrate how these  
 524 features can be institutionalized through multi-stakeholder governance (Carlson, 2025).

## 525 8 CONCLUSION

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

540 REFERENCES  
541

542 Regulation (EU) 2024/1689 of the European Parliament and of the Council establishing harmonised  
543 rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union,  
544 L 168/1, 2024a. URL <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689>. Adopted 13 June 2024.  
545

546 Directive (EU) 2024/2853 of the European Parliament and of the Council on liability for defective  
547 products and repealing Council Directive 85/374/EEC. Official Journal of the European Union, L  
548 2024/2853, 18 November 2024, 2024b. URL <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024L2853>. Adopted 23 October 2024; entry into force  
549 20 days after publication.  
550

551 Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K  
552 Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv*  
553 preprint *arXiv:2508.10925*, 2025.  
554

555 Mir Masood Ali, David G Balash, Monica Kodwani, Chris Kanich, and Adam J Aviv. Honesty is  
556 the best policy: On the accuracy of apple privacy labels compared to apps' privacy policies. *arXiv*  
557 preprint *arXiv:2306.17063*, 2023.  
558

559 American Law Institute. *Restatement (Third) of Agency*. American Law Institute, Philadelphia, PA,  
560 2006.  
561

562 Merve Astekin, Max Hort, and Leon Moonen. An exploratory study on how non-determinism in  
563 large language models affects log parsing. In *Proceedings of the ACM/IEEE 2nd International*  
564 *Workshop on Interpretability, Robustness, and Benchmarking in Neural Software Engineering*,  
565 pp. 13–18, 2024.  
566

567 Ian Ayres and Jack M Balkin. The law of ai is the law of risky agents without intentions. *U. Chi. L.*  
568 *Rev. Online*, pp. 1, 2024.  
569

570 Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones,  
571 Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harm-  
572 lessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.  
573

574 Jack M Balkin. Information fiduciaries and the first amendment. *UCDL Rev.*, 49:1183, 2015.  
575

576 Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the  
577 dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM*  
578 *conference on fairness, accountability, and transparency*, pp. 610–623, 2021.  
579

580 Yoshua Bengio, Sören Mindermann, Daniel Privitera, Rishi Besiroglu, Tamay andBommansani,  
581 Stephen Casper, Yejin Choi, Philip Fox, Ben Garfinkel, Hoda Goldfarb, Danielleand Heidari, An-  
582 son Ho, Sayash Kapoor, Leila Khalatbari, Shayne Longpre, Sam Manning, Vasilios Mavroudis,  
583 Mantas Mazeika, Julian Michael, Jessica Newman, Chinasa T. Ng, Kwan Yeeand Okolo, Deb-  
584 orah Raji, Girish Sastry, Elizabeth Seger, Tobin Skeadas, Theodora andSouth, Emma Strubell,  
585 Florian Tramèr, Lucia Velasco, Nicole Wheeler, Daron Acemoglu, Olubayo Adekanmbi, David  
586 Dalrymple, Thomas G. Dietterich, Pascale Felten, Edward W. andFung, Pierre-Olivier Gourin-  
587 chas, Fredrik Heintz, Geoffrey Hinton, Andreas Jennings, Nick andKrause, Susan Leavy, Percy  
588 Liang, Teresa Ludermir, Vidushi Marda, Helen Margetts, John McDermid, Jane Munga, Arvind  
589 Narayanan, Alondra Nelson, Alice Neppel, Clara andOh, Gopal Ramchurn, Stuart Russell, Ma-  
590 rietje Schaake, Dawn Schölkopf, Bernhard andSong, Alvaro Soto, Lee Tiedrich, Gaël Varo-  
591 quaux, Andrew Yao, Olubunmi Zhang, Ya-Qinand Ajala, Fahad Albalawi, Marwan Alserkal,  
592 Guillaume Avrin, André Carlos Ponce de Leon Ferreira Busch, Christian andde Carvalho, Bron-  
593 wyn Fox, Amandeep Singh Gill, Ahmet Halit Hatip, Juha Heikkilä, Chris Johnson, Gill Jolly, Ziv  
594 Katzir, Saif M. Khan, Hiroaki Kitano, Antonio Krüger, Kyoung Mu Lee, Dominic Vincent Ligot,  
595 José Ramón López Portillo, Oleksii Molchanovskyi, Andrea Monti, Nusu Mwamanzi, Nuria  
596 Nemer, Mona andOliver, Raquel Pezoa Rivera, Balaraman Ravindran, Hammam Riza, Ciarán  
597 Rugege, Crystal andSeoighe, Jerry Sheehan, Haroon Sheikh, Denise Wong, and Yi Zeng. Inter-  
598 national ai safety report. Technical Report DSIT 2025/001, 2025. URL <https://www.gov.uk/government/publications/international-ai-safety-report-2025>.  
599

594 Howard Bennett. *Principles of the Law of Agency*. Bloomsbury Publishing, 2024.  
 595

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

599 Leonard Bereska and Efstratios Gavves. Mechanistic interpretability for ai safety—a review. *arXiv*  
 600 *preprint arXiv:2404.14082*, 2024.  
 601

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

604 Bushey v. United States. *Bushey v. United States*, 1968. 398 F.2d 167 (2d Cir. 1968).  
 605

606 Tiago Sérgio Cabral. Liability and artificial intelligence in the eu: Assessing the adequacy of the  
 607 current product liability directive. *Maastricht Journal of European and Comparative Law*, 27(5):  
 608 615–635, 2020.

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

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

614

615 Inyoung Cheong. Epistemic and emotional harms of generative ai: Towards human-centered first  
 616 amendment. *Available at SSRN 5435335*, 2025.

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

619

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

621

622 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.

623

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

625

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

629

630 Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben  
 631 Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. Red teaming language models to  
 632 reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*,  
 633 2022.

634

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

638

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

640

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

644

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

646

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

648 Claudia E Haupt. Platforms as trustees: information fiduciaries and the value of analogy. *Harv. L.  
649 Rev. F.*, 134:34, 2020.  
650

651 Haybeck v. Prodigy Servs. Co. Haybeck v. Prodigy Servs. Co., 1996. 944 F. Supp. 326 (S.D.N.Y.  
652 1996).

653 Luxi He, Nimra Nadeem, Michel Liao, Howard Chen, Danqi Chen, Mariano-Florentino Cuéllar,  
654 and Peter Henderson. Statutory construction and interpretation for artificial intelligence. *arXiv  
655 preprint arXiv:2509.01186*, 2025.  
656

657 Juyeon Heo, Miao Xiong, Christina Heinze-Deml, and Jaya Narain. Do llms estimate uncertainty  
658 well in instruction-following? *arXiv preprint arXiv:2410.14582*, 2024.  
659

660 Hinman v. Westinghouse Electric Co. Hinman v. Westinghouse Electric Co., 1970. 2 Cal.3d 956  
(1970).  
661

662 Laurie Hughes, Yogesh K Dwivedi, Tegwen Malik, Mazen Shawosh, Mousa Ahmed Albashrawi,  
663 Il Jeon, Vincent Dutot, Mandanna Appanderanda, Tom Crick, Rahul De', et al. Ai agents and  
664 agentic systems: A multi-expert analysis. *Journal of Computer Information Systems*, pp. 1–29,  
665 2025.  
666

667 Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang,  
668 Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM  
computing surveys*, 55(12):1–38, 2023.  
669

670 Margot E Kaminski. Regulating the risks of ai. *BUL Rev.*, 103:1347, 2023.  
671

672 Sayash Kapoor, Noam Kolt, and Seth Lazar. Build agent advocates, not platform agents. *arXiv  
preprint arXiv:2505.04345*, 2025.  
673

674 Lina M Khan and David E Pozen. A skeptical view of information fiduciaries. *Harvard Law Review*,  
675 133(2):497–541, 2019.  
676

677 Leonie Koessler. Fiduciary requirements for virtual assistants. *Ethics and Information Technology*,  
678 26(2):21, 2024.  
679

680 Noam Kolt. Governing ai agents. *arXiv preprint arXiv:2501.07913*, 2025.  
681

682 William M Landes and Richard A Posner. *The economic structure of tort law*. Harvard University  
Press, 1987.  
683

684 Micheal Lanham. *AI agents in action*. Simon and Schuster, 2025.  
685

686 Lee v. United States. Lee v. United States, 2001. 171 F.Supp.2d 566, 575–577 (M.D.N.C. 2001).  
687

688 Anat Lior. Ai entities as ai agents: Artificial intelligence liability and the ai respondeat superior  
analogy. *Mitchell Hamline L. Rev.*, 46:1043, 2019.  
689

690 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.  
691

692 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  
693 Empirical Legal Studies*, 22(2):216–242, 2025.  
694

695 Monty v. Or-landi. Monty v. Or-landi, 1959. 337 P.2d 861, 865–866 (Cal. App. 1959).  
696

697 Norman Mu, Sarah Chen, Zifan Wang, Sizhe Chen, David Karamardian, Lulwa Aljeraisy, Basel  
698 Alomair, Dan Hendrycks, and David Wagner. Can llms follow simple rules? *arXiv preprint  
arXiv:2311.04235*, 2023.  
699

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

702 Roderick Munday. *Agency: Law and principles*. Oxford University Press, USA, 2010.  
 703

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

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

710 Cullen O’Keefe, Ketan Ramakrishnan, Janna Tay, and Christoph Winter. Law-following ai: Design-  
 711 ing ai agents to obey human laws. 2025.

712 OpenAI. Introducing chatgpt pulse. [https://openai.com/index/introducing-c  
 713 hatgpt-pulse/?ref=platformer.news](https://openai.com/index/introducing-chatgpt-pulse/?ref=platformer.news), 2025. Introducing a Pulse that proactively  
 714 initiates asynchronous research to deliver personalized updates based on users’ chats, feedback,  
 715 and connected apps like calendar and email.

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

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

719 Neil Richards and Woodrow Hartzog. A duty of loyalty for privacy law. *Wash. UL Rev.*, 99:961,  
 720 2021.

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

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

724 Walid S Saba. Stochastic llms do not understand language: towards symbolic, explainable and  
 725 ontologically based llms. In *International conference on conceptual modeling*, pp. 3–19. Springer,  
 726 2023.

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

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

731 Brian T Smith. Sending agents to the principal’s office: How talent agency packaging and producing  
 732 breach the fiduciary duties agents owe their artist-clients. *UCLA Ent. L. Rev.*, 27:173, 2019.

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

737 Diana Stern and Dazza Greenwood. From fine print to machine code: How ai agents are rewriting  
 738 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.

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

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

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

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

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

759  
 760 Rui-Jie Yew, Bill Marino, and Suresh Venkatasubramanian. Red teaming ai policy: A taxonomy of  
 761 avoision and the eu ai act. In *Proceedings of the 2025 ACM Conference on Fairness, Account-*  
 762 *ability, and Transparency*, pp. 404–415, 2025.

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

764  
 765 Shenzhe Zhu, Jiao Sun, Yi Nian, Tobin South, Alex Pentland, and Jiaxin Pei. The automated but  
 766 risky game: Modeling agent-to-agent negotiations and transactions in consumer markets. *arXiv*  
 767 *preprint arXiv:2506.00073*, 2025.

## 768 769 A APPENDIX

### 770 A.1 USE OF LARGE LANGUAGE MODELS

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

### 773 A.2 LEGAL DISCUSSIONS ON AGENCY IN DIGITAL SERVICES

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

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

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

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

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

### 802 A.3 DELEGATION, INTERACTIVITY, AND AUTONOMY ON DIGITAL SERVICES

### 803 A.4 COURTS’ REASONING ON RESPONDEAT SUPERIOR

810  
811  
812  
813  
814  
815

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

817 <b>Property</b>	818 <b>Wikipedia</b>	819 <b>Amazon</b>	820 <b>AI agents</b>	821 <b>Human Agents</b>
<b>822 Delegation</b>	823 Users retrieve information directly; no task execution.	824 Users specify items and transactions; platform executes predefined workflows.	825 Users delegate goals (“book me a flight”); agent decomposes into subtasks, applies constraints, executes.	826 Users delegate outcomes broadly; human agent interprets intent, applies judgment, handles exceptions.
<b>827 Interactivity</b>	828 Static interaction: query and read results; no context across sessions.	829 Structured interactions: browse, purchase, track; limited conversational support.	830 Dynamic, multimodal: natural language conversations, API calls, negotiation with third parties, memory of context.	831 Rich, adaptive: nuanced communication, persuasion, empathy, social intelligence.
<b>832 Autonomy</b>	833 None: system is passive, user-driven.	834 Low: limited automation (recommendations, order tracking) but not proactive.	835 Medium-High: initiative (event triggers), adaptation (plan revision), persistence (long-running workflows).	836 High: can self-initiate, deeply adapt, sustain long-term projects, improvise under uncertainty.

837  
838  
839  
840  
841  
842

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

844 <b>Employee Conduct</b>	845 <b>Employer Liable?</b>	846 <b>Rationale</b>
847 Employee makes intentional misrepresentations to prospective customers to induce purchases	848 Yes	849 Making statements to customers is within assigned job duties <i>Quick v. Peoples Bank</i> (1993)
850 Employee drives negligently while performing delivery duties	851 Yes	852 Driving is part of assigned task; negligence is foreseeable <i>Hinman v. Westinghouse Electric Co.</i> (1970)
853 Employee slams trays during heated customer complaint, injuring customer	854 Yes	855 Emotionally-driven conduct while performing assigned work (handling complaints) <i>Lee v. United States</i> (2001)
856 Truck driver chats on cell phone, becomes distracted, and causes accident	857 No	858 Personal phone call is a non-work-related independent course of action <i>Haybeck v. Prodigy Servs. Co.</i> (1996)
859 Irate driver shoots another driver while driving company truck	860 No	861 Extreme violence exceeds any reasonable scope of employment <i>Monty v. Or-landi</i> (1959)
862 Inebriated seaman turns valves on drydock wall, causing flooding and ship damage	863 Yes	864 Foreseeable risk of seamen’s conduct; act not entirely due to personal life <i>Bushey v. United States</i> (1968)

865  
866  
867  
868  
869