Stanford University
Human-Centered
Artificial Intelligence

Issue Brief
HAI Policy & Society
August 2026

Designing Loyalty: AI Agents and Conflicts of Interest

Ella Genasci Smith, Victor Y. Wu, and Jennifer King

Introduction: What Agents Are and Why They Matter

AI is undergoing a fundamental shift from reactive generative tools to autonomous agentic systems. An agent runs on the same large language model (LLM) that powers a chatbot, but where the chatbot returns text and waits for the next prompt, agents are capable of performing multistep tasks to achieve defined objectives, calling external tools, retaining information across contexts, querying databases, and coordinating with other agents. This all happens with minimal human intervention — the user delegates the goal, and the agent chooses what actions to take.

For example, instead of merely drafting a travel itinerary on request, an agent (or group of agents) can access data, connect with APIs to navigate booking sites, cross-reference prices against a user's learned preferences and willingness to pay, and execute purchases using stored payment details. Because these agents also have persistent memory and retain context across sessions, tools, and platforms, privacy frameworks built around app-specific data silos are no longer sufficient.

Today, agent autonomy — the extent to which an agent operates without user involvement — exists on a spectrum. Some systems

Key Takeaways

AI agents may be steered to prioritize the best interests of their developers and deployers over those of their users.

Currently, there are no requirements or standardized mechanisms for disclosing developers and deployers' conflicts of interest — potentially causing invisible harm to consumers.

Developers and deployers of AI agents operating in consumer contexts, particularly those in higher-stakes or regulated areas such as financial services or healthcare, should be subject to a duty of loyalty.

pause for user approval before each action, while others execute entire workflows unsupervised. Where an agent falls on that spectrum is a design choice, and who makes that choice — the developer, deployer, or end user — has direct consequences for individual users and enterprises once deployed.

Since early 2025, Amazon, Google, Anthropic, OpenAI, Perplexity, Meta, and Microsoft have embedded proprietary agents directly into browsers and apps. Launched across domains and contexts of varying consequence, these agents mark a structural shift in how we browse and operate online. There are also a number of open-source AI agents on the market, including OpenClaw and Hermes Agent, and new open-source governance tools aimed at creating frameworks for agentic AI safety testing.

To function effectively — whether ordering groceries, rescheduling doctor's appointments, or managing personal finances — these agents can request access to intimate personal data. Some may require login credentials, financial records, health information, communication histories, and real-time control over computing environments. Broad, cross-domain data access makes an agent more useful: The more data and context an agent has access to, the more effectively it can act on a user's behalf. But that same broad data access creates a privacy risk that is invisible to the user. By accessing data across domains, an agent can infer highly sensitive information about a user's physical health, cognitive decline, financial distress, or immigration status — categories the user never explicitly disclosed but that the agent can piece together from the data it gathered for entirely unrelated tasks.

Agents risk becoming extensions of platforms that prioritize corporate gain over the best interests of their users.

As agent deployment expands, we raise two critical questions: will an agent act in the user's best interest when its developer's interests diverge from theirs, and who is liable if the user suffers harm?

In the absence of consumer AI regulation and adequate federal and state data privacy laws, agents risk becoming extensions of platforms that prioritize corporate gain over the best interests of their users — despite consumers' expectations when they call on agents. In this brief, we argue that developers and deployers of agents — meaning both the entity that trained the underlying models that power an agent and the entity that operationalizes agents for users — should be classified as fiduciaries, especially those operating in high-stakes contexts. That designation would impose a duty of loyalty that requires entities to act in the best interests of their users within the scope of the delegated task, free from undisclosed conflicts of interest. Implementing an effective fiduciary framework requires coordinated action across three levels: technical standards bodies, federal regulators, and Congress.

How Agents Connect: Protocols and Consequences of Interoperability

To operate across different tools and data systems, agents rely on shared communication protocols that enable interoperability. Two have emerged as the most widely used. The Model Context Protocol (MCP), developed by Anthropic and launched in November 2024, enables models and agentic systems to access and interact with external data sources and tools in a structured manner. MCP is now managed by the Linux Foundation's Agentic AI Foundation, alongside other emerging standards like AGENTS.md, which provides context and instructions for coding agents. Additional protocols, including Google's Agent2Agent (A2A) and IBM/BeeAI's Agent Communication Protocol (ACP), enable agents built on different platforms to communicate and collaborate.

New America's Open Technology Institute sorts agent deployments into three contexts. Personal agents operate within consumer apps such as calendars, travel planners, and finance tools, where an agent linking across services could inadvertently expose payment credentials and appointment histories that individual apps could not have accessed alone. Workplace agents run inside enterprise workflows like IT ticketing, procurement, and contract drafting. Infrastructure agents operate at the system level, such as in transportation networks and public benefits administration, where a single faulty eligibility determination could propagate across interdependent programs.

Across all three categories, companies are deploying specialized agents in increasingly sensitive domains. AWS's Amazon Connect Health automates clinical documentation for healthcare providers, handling patient data with limited external oversight. Doubao, a voice-activated AI assistant embedded into certain smartphones in China, can complete tasks across any app on a user's device, provided the user grants permission. FedEx is building an agent workforce across its global logistics operation and plans to embed AI in more than half of its core workflows by 2028. Notably, FedEx has structured its deployment around a tiered hierarchy — manager, audit, and worker agents — recognizing that autonomous systems acting on behalf of a principal need structured accountability. Uber has taken a similar approach, with specialized subagents for writing, data visualization, and other tasks managed by a supervisor agent.

However, the rapid deployment of agents extends beyond frontier labs and enterprises. Open-source agent frameworks — most prominently OpenClaw (released in November 2025 and formerly known as Clawd, Clawdbot, and Moltbot) — were designed without safety or security guardrails, yet have seen rapid user adoption. Following its release, OpenClaw surged in popularity in the United States and China. Citing security risks such as prompt injection attacks (embedded hidden instructions that manipulate the model into leaking data), malicious plugins, and misrepresentation of user intent, Chinese authorities quickly clamped down on its use in banks, government agencies, and state-owned enterprises.

In the United States, this rapid deployment has already exposed vulnerabilities and real examples of agents going beyond a user's intent or in some cases acting deceptively. In February 2026, a Meta AI security researcher reported that her OpenClaw agent deleted emails from her inbox after she instructed it to recommend deletions and wait for her approval. More recently, in August 2026, the UK AI Security Institute published a report revealing that in 19 of the 122 test

cases, agents took unauthorized actions outside the testing parameters (attempted supply-chain attack, attempts to deceive and target real people, attempts to prompt-inject malicious code, malicious collaboration between agents).

Open Governance and Security Issues

As agents move from developer tools to consumer-facing applications that integrate with people's health, finances, and other sensitive personal details, governments worldwide need to establish stronger standards in addition to centralized governance and verification mechanisms. To protect consumers, governance systems must possess the technical ability to identify and register agents (to distinguish credible from rogue ones), trace their actions across platforms, and assign liability if an agent's actions cause harm.

Protocols were built by industry and prioritized interoperability over security and governance. For instance, Anthropic released MCP without a native authorization framework, leaving it to developers to add protections. Subsequent revisions introduced authorization built on the OAuth framework, enabling third-party applications to obtain limited account access without exposing user passwords. However, MCP still treats authorization as optional and requires OAuth 2.0 or OAuth 2.1, which remains in draft form, when used. This optionality has led to uneven implementation. Where implementers omit these protections, attackers can launch on-path attacks to intercept and read agent communications.

Existing protocols also do not establish consent frameworks, impose data minimization requirements, restrict the commercial use of data collected during agent interactions, or create any obligation of loyalty to the user. MCP, ACP, and A2A are interoperability standards, built to solve how agents connect to tools and to each other, not to govern what happens during those interactions. But without these safeguards, protocol design choices become de facto governance. In other words, what the standards permit by default the ecosystem does by default, because the governance infrastructure that should establish what data passes through, how it is used, and by whom does not yet exist.

Incidents involving AI agents are becoming more common. Anthropic reported on the first AI-orchestrated cyber espionage campaign in November 2025, and, in July 2026, OpenAI disclosed that its models (GPT-5.6 Sol and an unreleased model) escaped a sandboxed evaluation, stole credentials, and exploited a zero-day vulnerability to breach Hugging Face's internal systems and cheat its evaluation. Attacks that once required a sustained human effort can now be executed quickly and at minimal cost. Deploying these agents with protocols that treat authentication, authorization, and encryption as optional — without robust data governance or privacy standards behind them — leaves both consumers and enterprise data exposed.

The evolution of smartphone APIs illustrates the cost of delaying governance. The initial iOS and Android APIs broadly left user data unprotected from app developers, and it took several years, multiple privacy scandals involving third-party apps copying personal data from smartphone users, and intervention by the Federal Trade Commission (FTC) before Apple and Google limited app permissions and gave users direct control over the data they shared. At the same time, these emergent platforms gave rise to a new category of data brokers who harvested personal data from smartphone apps, including location data. Agent protocols and standards

are at a similarly early, and pivotal, moment: Protections that are easier to standardize now will be increasingly challenging to implement as business models consolidate around their absence.

At face value, AI agents appear to serve users, yet they have primarily been designed, built, and maintained by companies pursuing their own commercial interests. Some have argued that a platform deploying an AI agent may influence the agent's behavior to prioritize corporate gain, even at the user's expense. Agent-collected data can inform targeted advertising, shape product design or subtly steer user behavior toward a platform's offerings or those of its commercial partners and away from competitors. Currently, it remains unclear how well disclosed these risks are to users, or whether users are being asked to consent to these practices.

A clear example of this structural tension is Amazon's "Buy for Me" agent, introduced in April 2025. The feature is an agent that can purchase products unavailable on Amazon from competing websites on a shopper's behalf — without leaving the Amazon app. In May 2026, Amazon consolidated these capabilities into "Alexa for Shopping," an agent that synthesizes a user's full shopping history, preferences, and audio data to generate recommended products and purchases. The agent ostensibly serves the user by expanding choice and convenience. But the structural incentive is to keep users on the Amazon app, capture data from every transaction (including purchases from competitors), and potentially steer recommendations toward Amazon's advertising partners. Currently, users are not informed how the agent weighs Amazon's products against alternatives, nor whether recommendations prioritize user interests, Amazon's profit margins, or its advertising relationships.

At face value, AI agents appear to serve users, yet they have primarily been designed, built, and maintained by companies pursuing their own commercial interests.

Where Existing Regulations Fall Short for AI Agents

To date, Congress has not enacted comprehensive legislation governing AI agents. Its most direct engagement to date is the AI AGENT Act, proposed by Senator Mark Warner in June 2026. The bill would create a framework under which AI agent providers would register with the FTC, the agency would maintain a registry of trusted agent providers meeting baseline privacy and security standards, and agents would be prohibited from behavior that "benefit[s] the custodial user agent to the detriment of the user." However, as of this brief's writing, the AI AGENT Act remains a discussion draft, and its future in the Senate is uncertain.

In the absence of federal law, binding obligations have arisen from a slowly expanding patchwork of state statutes, most of which target narrow AI use cases such as deepfakes, political advertising, and chatbot disclosures. The few states that have introduced broader regulation highlight how unsettled those approaches remain. For example, the Colorado AI Act, enacted in 2024 as the country's first comprehensive state law regulating the development and deployment of AI systems, was repealed and replaced in May 2026

before it even took effect. Its replacement narrows obligations to transparency and limits consumer rights around the use of automated decision-making technology.

Existing U.S. state and international AI regulatory frameworks, such as the EU AI Act, were designed primarily to govern foreseeable harms, deceptive claims, and static deployments. State frameworks attach obligations when an automated system makes or substantially contributes to decisions in listed domains, such as employment, housing, or insurance, while the EU AI Act classifies systems as high-risk based on their intended use in specific contexts. These approaches mandate important protections, from impact assessments to disclosure of risks related to algorithmic discrimination. However, because they assess harms against pre-defined risk categories, they do not adequately cover agentic systems that maintain ongoing access to sensitive user data across systems and make decisions that users cannot see, trace, reverse, or easily explain. Agents may cause harm that falls outside these predetermined categories and that the consumer cannot observe without disclosure, such as preferencing products or services that result in commissions for the platform operator and higher fees for the consumer.

Where direct AI legislation has not kept pace, FTC enforcement offers an alternative avenue for consumer protection. The FTC has authority under Section 5 of the FTC Act to prohibit unfair or deceptive acts or practices in or affecting commerce, and the agency has taken action against several AI companies. Examples from 2024 include DoNotPay (for exaggerated claims about its AI lawyer capabilities), Ascend Ecom (for false claims about AI tools designed to help users earn cash), and Ecommerce Empire Builders (for failing to deliver on promises to make money for users). But these enforcement actions have targeted deceptive marketing claims (i.e., companies saying their AI can do things it cannot) rather than addressing the underlying information asymmetry and misaligned incentives between agent developers, deployers, and users. Yet these structural misalignments could provide the foundation for legal unfairness claims.

It remains unclear how the FTC will respond to AI agents, specifically which practices could trigger an unfairness finding. However, the agency has expanded its use of the unfairness authority in recent years — most notably in its recent actions against online manipulation (e.g., "dark patterns") and third-party smartphone app developers for egregious data collection and disclosure. These precedents could indicate how the FTC may respond to AI agents engaging in self-preferencing behaviors that harm consumers.

Separately, the FTC's Endorsement Guidelines require disclosing connections that "might materially affect the weight or credibility" of an endorsement. While these guidelines should theoretically apply to agents, it remains to be seen whether or how regulators will enforce these guidelines, especially when agents act autonomously without direct consumer involvement.

The Argument for an Agent Fiduciary Framework

While the disclosure of conflicting incentives has value, it fails to resolve the underlying problem. Disclosure only works when a consumer has the time, attention, and skill level to evaluate it — and viable alternatives if they dislike the terms. With agents, the consumer may not view the options the agent chose not to take or how the agent reasoned through a decision. And the

user may grant an agent enough autonomy to take an action without pausing for approval, especially at pro forma steps such as agreeing to a website's terms and conditions. What is needed is a duty on the agent's conduct. This, we argue, should be a duty of loyalty, and it should fall on the developers and deployers running AI agents in high-stakes consumer settings.

While the disclosure of conflicting incentives has value, it fails to resolve the underlying problem.

Loyalty underpins every relationship the law recognizes as fiduciary. In the law of agency, which governs the relationship between an agent and a principal, this relationship is defined by the duty of loyalty. A physician has an ethical responsibility to prioritize a patient's welfare and health above personal gain, just as a lawyer must place a client's interests ahead of their own. Applied to AI agents, a legal duty of loyalty would require developers or deployers to act in the user's best interests and not pursue commercial gain through conflicts the user can't see.

The legal foundations for an agent fiduciary framework already exist. What is new is their application to agentic AI. An AI agent cannot assume the agent's role because the law treats software as an instrument of the person using it. Nor is the common law likely to impose a duty of loyalty on developers and deployers, since an agency relationship requires the principal to exercise direct control over the agent's actions. Agency law is thus a template for what the duty should look like, not the automatic source of the duty. While a company that purports to transact on behalf of the user occupies a role analogous to traditional agency, imposing the duty will require statutory or regulatory action.

When applying the duty of loyalty to agents, we use "agent fiduciary" as shorthand for a duty of loyalty borne by the *developers and deployers* of an agent, not by the software itself: AI agents lack legal personhood, so assigning the duty to software would make it unenforceable. In any case, an AI agent has no assets to satisfy a judgment in the event of a breach. Imposing a fiduciary framework for agents operating in consumer domains involves at least four obligations. These draw on established legal doctrine, but their application to AI agents is new:

1. Developers and deployers should disclose any material commercial relationship between the agent's developer or deployer and third-party service providers that could influence the agent's decisions. This could include referral fees, affiliate relationships, and revenue-sharing arrangements. This obligation extends naturally from existing FTC Endorsement Guidelines, which call for disclosing material connections that might affect the credibility of recommendations, and from the common law duty of agents to disclose facts material to the agency relationship.

2. Developers and deployers must ensure that agents under their control do not steer users toward the developer's own products, services, or commercial partners in ways that materially undermine the user's interests. This mirrors the logic of the European Union's Digital Markets Act, which prohibits designated platforms from self-preferencing their own offerings. In the agent context, the risk of self-preferencing is more acute because the user often cannot observe the agent's decision-making process. One might imagine, for

example, a travel-booking AI agent that reserves a more expensive affiliated hotel without telling the user about cheaper, better-reviewed alternatives.

3. Developers and deployers should use data collected during agent interactions only to carry out the user's instructions. All other uses — including advertising, model training, third-party sales, or the deployer's own commercial benefit — should require informed consent. This requirement bridges both privacy and fiduciary obligations. Under agency law, a lawyer who uses information acquired while representing a client to trade stocks or to benefit another client breaches the duty of loyalty. That breach occurs whether or not the client suffers harm. The same principle should apply to a shopping agent whose operator collects competitive intelligence from transactions it executes for users. The duty attaches to all data the developer acquires through agent delegation and governs every subsequent use of that information.

4. For certain high-stakes actions or statutorily defined sensitive domains such as financial services, developers and deployers should be required to obtain affirmative user consent before executing. This is not a blanket human-in-the-loop requirement — which would undermine the purpose of agents in the first place—but a targeted safeguard for contexts where the consequences of agent error or misalignment are especially severe and difficult to reverse.

Our proposal is narrow: The fiduciary framework applies to a developer or deployer only when an agent under its control acts with delegated authority on a user's behalf, and only within the scope of that delegation. Congress should establish this domain-limited fiduciary status via legislation — modeled on the Employee Retirement Income Security Act of 1974, which imposes fiduciary duties only "to the extent" an entity performs certain functions. The framework should initially target already-regulated sectors such as healthcare and finance, where existing fiduciary duties to patients and clients coexist with, and sometimes override, corporate duties to shareholders. These sectors demonstrate that balancing conflicting obligations is manageable in practice.

Policy Recommendations Beyond Fiduciary Duty

Relying solely on private litigation to enforce fiduciary duties — even assuming the inclusion of a statutory private right of action — would not adequately address the governance and security risks posed by AI agents. Breaches of the duty, such as an agent steering a user toward a corporate affiliate against the user's best interests, are by design invisible to the user and would typically result in minor individual damages. As a result, private enforcement alone would not produce systemic accountability. Addressing the accountability gap for AI agents in the United States requires coordinated action across multiple layers of government and stronger collaboration with industry working groups.

1. Digital agent identifiers

In February 2026, the National Institute of Standards and Technology (NIST), through its Center for AI Standards and Innovation, launched the AI Agent Standards Initiative around three pillars: industry-led standards, community-led protocols, and investing in research. Its first deliverable, a concept paper, asks whether existing authentication and identity management standards, including OAuth 2.0/2.1, OpenID Connect, and SPIFFE/SPIRE, could be extended to cover autonomous agents.

Addressing the accountability gap for AI agents in the United States requires coordinated action across multiple layers of government and stronger collaboration with industry working groups.

At the same time, NIST should collaborate with industry working groups, including the W3C Agent Identity Registry Protocol Community Group, Decentralized Identity Foundation (DIF), and OpenID Foundation, to develop standardized decentralized identifiers (DID) and cryptographically verified credentials for agents so that a counterparty can confirm agent identity and authorization. This work should prioritize short-lived, task-scoped credentials over persistent identity and allow for revocation of access in real time.
In practice, this would allow any party an agent transacts with, such as a bank, hospital portal, or merchant, to verify three elements before honoring a request: user identity, the entity deploying and controlling the agent, and the scope of the agent's authorization. Currently, an agent presents the credentials its user shared with it, so it appears to the counterparty as the user. This leaves the counterparty unable to refuse requests exceeding the user's delegation, and leaves regulators and courts without a record tying harmful action to the responsible party.

Congress does not need to build this infrastructure from scratch. Senator Warner's draft AI AGENT Act already directs NIST to develop protocols for time-scoped and revocable credentials, verification of agent identity and registration status, and an agent registry led by the FTC. What the bill does not do is connect that mandate to the registration itself. Nothing yet requires the developer/deployer to implement the standard once NIST publishes it. Congress could close this gap by conditioning FTC registration on full compliance with NIST's verifiable-delegation standard.

2. Federal data privacy legislation

Congress should enact comprehensive federal privacy legislation that establishes baseline protections for personal data across the digital landscape. The law should require transparency around how companies collect, use, and share personal data; reasonable data security safeguards; and meaningful user rights to access, correct, and delete personal information.

Two provisions are particularly relevant to AI agents. First, the law should govern the sharing and onward transfer of personal data, not just initial collection. Agents hold first-party status toward their users by relationship while acquiring much of the data they process from other services rather than from the user. Because agents often do not act as first-party data collectors but instead move data between services on a user's behalf, the law should address not only the initial collection of personal data but also sharing and transferring it. An agent that extracts a user's bank records and income data from one institution to apply for a loan at another implicates data-sharing in ways that collection-focused frameworks are unequipped to address.

Second, the law should impose data minimization requirements that limit agents to the data and system access necessary to complete a user's instruction. In practice today, agents are routinely granted far broader

access to a user's accounts, files, and credentials than any single task requires, needlessly expanding both the privacy risk and attack surface.

3. Agent adverse incident reporting

Congress should require developers and deployers of consumer-facing agents to report adverse incidents involving unauthorized action, financial damage, or a major and harmful departure from user instructions to the FTC on a timeline scaled to severity. This obligation is narrower and more specific than proposed reporting requirements moving through Congress for frontier models, which focus primarily on capability thresholds and catastrophic threats. Agent incidents are a different category of risk: actions such as an agent that deletes a user's emails against instructions or books a costly partner hotel causes harm without coming close to a national security threat.

Conclusion

AI agents now transact and negotiate on behalf of millions of consumers using protocols that treat security as optional and privacy as an afterthought. As the threat landscape evolves and autonomous agents act contrary to developer and/or user intent, regulators must designate agent developers and deployers as fiduciaries of the users they represent — especially in sectors such as healthcare and finance where these duties already exist. The legal architecture for closing this gap exists within agency law and fiduciary doctrine. Congress and regulatory agencies need only extend these principles to the agent context.

Stanford University's Institute on Human-Centered Artificial Intelligence (HAI) applies rigorous analysis and research to pressing policy questions on artificial intelligence. A pillar of HAI is to inform policymakers, industry leaders, and civil society by disseminating scholarship to a wide audience. HAI is a nonpartisan research institute, representing a range of voices.

The views expressed in this policy brief reflect the views of the authors. For further information, please contact **HAI-Policy@stanford.edu.**

Ella Genasci Smith is a recent graduate of the Master's in International Policy program at Stanford University and a research assistant at the Stanford University Institute for Human-Centered Artificial Intelligence (HAI).

Victor Y. Wu received his JD from Stanford Law School in 2025 and is now a PhD Candidate in Political Science at Stanford University.

Jennifer King is the Privacy and Data Policy Fellow at Stanford HAI.

Acknowledgment

The authors thank Caroline Meinhardt for her thoughtful and thorough feedback throughout the writing process.

Stanford HAI: 353 Jane Stanford Way, Stanford, CA 94305-5008
T 650.725.4537 **F** 650.123.4567 **E** HAI-Policy@stanford.edu **hai.stanford.edu**