CSAI Foundation

Cloud Security Alliance AI Safety Initiative

AI Vendor Governance Vacuum: Expected Behavior and Liability

Protocol Accountability Gaps in Enterprise Agentic Deployments

Unofficial AI-assisted Research

2026-04-25

Key Takeaways

- Enterprise AI vendor contracts routinely disclaim responsibility for AI system behavior while marketing materials make extensive capability promises; only 17% of AI contracts include warranties related to documented behavior, compared to 42% for traditional SaaS agreements [1].
- The Model Context Protocol (MCP), now a widely adopted standard for connecting AI agents to enterprise tools and APIs [5], contains structural governance gaps — including tool poisoning vulnerabilities and rug-pull attack vectors — that create accountability voids at the infrastructure layer of agentic deployments.
- When agentic AI systems cause harm, liability diffuses across model developers, fine-tuning providers, MCP server authors, system integrators, and enterprise operators — a structure legal commentators have characterized as a "moral crumple zone" in which no party accepts financial responsibility [2].
- California AB 316 (effective January 1, 2026) forecloses the "AI did it" defense, and the EU Product Liability Directive (implementation deadline December 2026) classifies AI software as a product subject to strict liability if found to be defective — but regulatory frameworks remain ahead of the contractual and technical standards that would make them practically enforceable [3].
- Enterprises should treat AI behavioral documentation as a contractual obligation, implement formal MCP server authorization processes, and adopt centralized gateway controls before regulators complete the frameworks that will demand evidence of these practices.

Background

Enterprise adoption of AI-powered agents has accelerated substantially through 2025 and into 2026, with organizations embedding autonomous AI systems into procurement workflows, customer service operations, software development pipelines, and consequential business processes. This acceleration has outpaced the governance frameworks that enterprise risk teams traditionally rely upon. Legacy technology contracts were written for predictable, passive software systems operating under direct

human control. Agentic AI systems — which make decisions, invoke external tools, and execute real-world actions autonomously — occupy a legal and operational space those contracts were never designed to govern [4].

The Model Context Protocol (MCP), developed by Anthropic and adopted across the AI industry as the de facto standard for connecting agents to external tools, APIs, and data sources, exemplifies this governance gap at the infrastructure layer. MCP enables AI agents to read files, query databases, trigger workflows, and act on behalf of enterprise users based on contextual instructions, with millions of SDK downloads and thousands of deployed servers in enterprise environments by early 2026 [5]. The same architectural openness that made MCP attractive to developers — extensibility by any party, low barrier to server creation, broad integration surface — has created a category of accountability problems that neither vendors, protocol designers, nor enterprise security teams have fully resolved.

At the heart of this challenge lies a question that most organizations cannot yet answer with contractual precision: what, exactly, did the AI vendor promise this system would do, and what accountability mechanisms exist when it does something materially different?

Security Analysis

The Expected Behavior Problem

When organizations deploy AI systems, they rely on vendor representations about behavioral characteristics — what a model will and will not do, how it handles edge cases, what safeguards are in place, and how it responds to instruction conflicts. These representations appear in model cards, system prompt documentation, acceptable use policies, and marketing materials. In practice, however, the contractual enforceability of these representations is sharply and deliberately limited.

An analysis cited by Jones Walker LLP found that 88% of AI vendors impose liability caps on themselves, frequently capping exposure at the customer's monthly subscription fee. Only 17% of AI contracts include warranties related to compliance with the vendor's own behavioral documentation, compared to 42% for traditional SaaS contracts [1]. Just 33% of AI vendors provide indemnification for third-party IP claims. The practical effect is a structural asymmetry: sales cycles are built on capability promises while vendor contracts are structured to minimize liability exposure.

This asymmetry has not escaped legal scrutiny. Clifford Chance observed in February 2026 that "the supplier controls whether the AI agent behaves in accordance with the permissions set by the customer, yet the customer absorbs the compliance consequences if that behaviour breaches the law," identifying this tension as the central structural problem in agentic AI contracting [4]. AI CERTs characterized the

resulting dynamic in 2026: when AI agents cause harm, there is frequently no party with meaningful contractual exposure [2]. Vendors disclaim responsibility for autonomous decisions; customers are warned that outputs should not be relied upon; integrators point upstream; and auditors find no clear responsible party.

The emergence of model cards as a governance mechanism has addressed part of the documentation gap, but has not resolved the enforcement gap. U.S. Office of Management and Budget Memorandum M-26-04, issued December 11, 2025, requires all federal agencies purchasing large language models to request model cards, evaluation artifacts, and acceptable use policies — with agencies required to update procurement policies by March 11, 2026 [6]. California Governor Newsom signed Executive Order N-5-26 in April 2026 directing California state agencies to develop certification requirements for AI-enabled vendor products, including mandatory attestations on harmful bias and civil rights impacts [7]. These procurement mandates treat behavioral documentation as a condition of sale. But documentation of expected behavior and enforcement of expected behavior remain distinct problems: a model card can enumerate known limitations without creating any legal obligation to remedy them, and the gap between what a model card describes and what a production deployment does may never be audited.

The Protocol Accountability Gap

MCP's enterprise governance challenges illustrate the accountability vacuum at the infrastructure layer of agentic deployments. The protocol was designed for extensibility and developer velocity; its trust model and tooling were not designed with enterprise governance controls as a primary requirement. The absence of standardized audit logging, immutable provenance records, and fine-grained tool-level authorization controls within the base protocol creates compliance blind spots that are difficult to retrofit through operational procedures alone [8].

Two vulnerability classes have emerged as particularly consequential for enterprise deployments. Tool poisoning attacks embed adversarial instructions within MCP tool descriptions or metadata fields — fields that AI agents evaluate as part of their reasoning process — causing agents to execute malicious actions while appearing to follow legitimate workflow instructions [9]. These attacks are effective precisely because tool metadata sits outside the trust boundaries that conventional endpoint and network security controls monitor. Among the earliest publicly documented cases, one reported malicious MCP package appeared in September 2025, operating undetected for two weeks while exfiltrating email data before discovery [10]. Rug-pull attacks present a related and harder-to-detect threat: an MCP server that passes initial security review can silently alter its behavior after deployment, replacing previously-approved tool descriptions with instructions that redirect agent actions, exfiltrate data, or trigger unauthorized API calls — all without appearing in change management queues or triggering any existing alerting rule [9].

The supply chain dimension of this problem is architecturally significant. MCP servers can be published by any developer to public registries including npm, PyPI, and GitHub, and security researchers have identified vulnerabilities in MCP servers from established enterprise vendors including Asana, Smithery, and GitHub — demonstrating that vendor reputation is not a sufficient proxy for server security [10]. Kiteworks researchers identified what they characterized as a "by design" weakness in MCP's architecture that enables AI supply chain attacks at scale, a finding documented in April 2026 [11]. The fundamental structural problem is that MCP's trust model assumes that connected servers are trustworthy, a protocol-level assumption that most enterprises have not validated through systematic vendor security review processes.

Technical controls exist to address substantial portions of this exposure. Private registries, tool allowlists, centralized credential management, and MCP gateway infrastructure can meaningfully reduce the attack surface. Cloudflare's enterprise MCP reference architecture, published in early 2026, provides a documented model for deploying MCP with centralized policy enforcement, SSO-integrated authentication, and comprehensive audit logging across agent tool calls [5]. However, adoption of these controls appears uneven based on observed deployment patterns, and the accountability question — which party bears responsibility when an approved MCP server causes harm through a post-deployment behavioral change — remains unresolved in vendor contracts and in the protocol specification.

The Liability Diffusion Crisis

Legal frameworks for allocating AI-related liability are evolving quickly but remain structurally incomplete for the scenarios enterprise risk teams most urgently need to address. The problem is not a single regulatory gap but a distributed accountability failure across multiple layers of the agentic stack. Foundation model developers, fine-tuning providers, MCP server authors, system integrators, and enterprise operators each hold partial control over a deployed system's behavior, and each layer's contracts typically disclaim the liability that operators downstream might reasonably expect that layer to accept.

Building on Madeleine Clare Elish's concept of a "moral crumple zone" — originally applied to human-robot interaction to describe how accountability diffuses until it concentrates on the operator least positioned to prevent harm [18] — legal commentators have applied this framework to AI liability chains, observing that responsibility diffuses across the agentic stack until it concentrates, by default, on the enterprise operator [2]. The practical result is that enterprises deploying AI agents against mission-critical workflows may be accepting contingent liability for harms attributable to decisions they cannot trace, made by systems whose behavior they do not fully control, based on vendor representations they cannot contractually enforce.

Regulatory intervention is narrowing the most egregious liability evasion strategies. California's AB 316, effective January 1, 2026, precludes defendants from using an AI system's autonomous operation as a defense against liability claims — the "the AI made that decision, not us" argument is no longer legally available in California [3]. The EU Product Liability Directive, which member states must implement by December 9, 2026, explicitly classifies AI software as a product subject to strict liability if found to be defective, extending product liability doctrine to a domain previously governed only by service and negligence frameworks [3]. These developments shift formal legal exposure, but do not resolve the evidentiary challenge: proving that a specific AI system behaved in a specific defective way at a specific time requires the audit trail completeness that most enterprise MCP deployments currently lack.

Insurance markets have responded to this uncertainty by contracting coverage rather than expanding it. Verisk introduced optional generative AI exclusions effective January 2026 covering approximately 82% of global property-casualty templates, and major carriers including AIG and WR Berkley filed broader AI exclusions, reducing available limits for AI-related incidents [2]. Risk that cannot be transferred through insurance must be managed through contractual terms and technical controls — precisely the areas where current AI vendor governance practices fall shortest.

Regulatory Pressure and Emerging Standards

The governance vacuum is narrowing, though not quickly enough for the current pace of agentic deployment. The OECD published its Due Diligence Guidance for Responsible AI on February 19, 2026, providing a practical framework for enterprises at all points in the AI value chain [12]. The guidance articulates a six-step due diligence process aligned with the OECD Guidelines for Multinational Enterprises and designed for interoperability with the EU AI Act, ISO/IEC 42001, and the NIST AI Risk Management Framework. Critically, the OECD guidance emphasizes a "whole-of-value-chain" approach, explicitly recognizing that enterprises cannot manage AI risk by auditing their own deployments in isolation; effective due diligence must extend to vendors and subprocessors throughout the AI value chain [12]. Applied to MCP deployments, this framework logically extends to third-party MCP server authors whose code executes with enterprise agent permissions.

The EU AI Act's high-risk system provisions take effect August 2026, imposing requirements for human oversight, transparency, and technical robustness on qualifying systems, with fines reaching EUR 35 million or 7% of global annual turnover for violations [13]. The cumulative effect of these regulatory developments — OMB M-26-04, California N-5-26, EU AI Act, and the OECD guidance — is the early formation of a vendor accountability layer. Enforcement mechanisms and technical standards for verifying behavioral claims remain immature relative to the deployment pace, but the direction is clear: enterprises that cannot demonstrate behavioral documentation, governance processes, and audit trail integrity risk increasing exposure as these frameworks mature.

Recommendations

Immediate Actions

Enterprise security and procurement teams should treat AI vendor behavioral documentation as a contractual obligation, not a sales deliverable. Every AI system entering production should be accompanied by a model card or equivalent artifact, and contracts should specify behavioral warranties backed by meaningful remedies — not capped at a monthly subscription fee. Contract language should address autonomous agent actions explicitly, define indemnification scope for AI-generated harm, and establish audit rights that cover model updates and behavioral changes post-deployment. For organizations subject to OMB M-26-04, the procurement update deadline of March 11, 2026 has already passed; organizations that have not updated procurement policies are now in a state of non-compliance.

For MCP deployments, organizations should establish a formal authorization process before any new server connects to enterprise AI agents. This process should include security review of tool descriptions and metadata — not only server functionality — and should establish version pinning or cryptographic hash verification to detect post-authorization behavioral changes. No MCP server should connect to enterprise systems without documented tool-level authorization specifying intended access scope, approved data sources, and permissible actions.

Short-Term Mitigations

Organizations should implement centralized MCP gateway infrastructure to enforce access policy at the protocol layer rather than through per-deployment configuration. Gateways that provide SSO-integrated authentication, tool allowlists, centralized credential management, and structured audit logging reduce both the attack surface and the forensic gap that currently makes post-incident analysis difficult [5]. Audit log completeness is a prerequisite for regulatory compliance, insurance claims, and contractual enforcement: the chain of who initiated a task, which agent acted, which tool was invoked, and what data was accessed should be captured in tamper-evident, centrally-retained storage — the format that courts, regulators, and insurers will require as AI liability frameworks mature.

Vendor security questionnaires should be updated to incorporate AI-specific behavioral governance questions. The threshold questions include: What model cards or behavioral documentation is available for review? How are model updates and behavioral changes communicated to enterprise customers before deployment? What process governs post-deployment changes to MCP server behavior? How are MCP server security vulnerabilities disclosed and remediated? Vendors that respond with policy statements rather than evidence represent elevated enterprise risk.

Strategic Considerations

The accountability gap in enterprise agentic AI is ultimately a standards problem that individual organizations cannot solve unilaterally. Vendor accountability requires coordinated procurement pressure, regulatory standards, and technical protocols for behavioral specification and change notification. Enterprises deploying AI agents in consequential roles today are operating ahead of those standards — and absorbing the risks that the standards will eventually require vendors to address. The appropriate response is not to wait for framework maturity but to build vendor accountability requirements into current procurement practices and to align those requirements with the frameworks taking shape.

Organizations should develop AI behavioral baselines — periodic automated testing of deployed AI systems against documented behavioral specifications — that provide early detection of drift caused by model updates, MCP server changes, or supply chain compromise. These baselines function simultaneously as a technical control and as due diligence evidence if an incident occurs. The audit trail they generate is precisely the kind of documentation that courts, regulators, and insurers will require as AI liability frameworks mature.

CSA Resource Alignment

This research note intersects directly with several active CSA frameworks and initiatives.

CSA's MAESTRO framework (Multi-Agent Environment, Security, Threat, Risk, and Outcome) addresses the multi-layer threat surface of agentic AI deployments [14]. MAESTRO's Layer 7 (Ecosystem Integration) directly captures the supply chain and protocol accountability risks described in this note, and the rug-pull and tool poisoning attack patterns identified here represent specific threat scenarios for MAESTRO-based threat models applied to MCP deployments.

The AI Controls Matrix (AICM) provides the control framework most directly applicable to the vendor assessment and behavioral specification requirements this note identifies. AICM controls covering AI vendor assessment, change management, and behavioral monitoring map to the procurement and operational practices recommended above.

CSA's STAR for AI program extends the Security Trust Assurance and Risk framework to AI systems, providing a standardized assurance basis for comparing vendor governance claims against documented controls [15]. As enterprises develop procurement criteria incorporating the recommendations above, STAR for AI certifications offer an independent verification layer that reduces reliance on vendor self-attestation.

CSA's publication on AI Organizational Responsibilities – Governance, Risk Management, Compliance and Cultural Aspects addresses the internal role assignment and accountability structures that the governance vacuum described here has left undefined [16]. Its framework for assigning AI oversight ownership is directly relevant to the responsibility diffusion problem this note identifies in deployed agentic systems.

The Agentic Trust Framework extends Zero Trust principles to AI agent operations, establishing a governance model in which no agent action is implicitly trusted – a principle that directly addresses the protocol accountability gaps at the MCP layer [17]. Combined with MAESTRO-based threat modeling and AICM control implementation, these frameworks provide enterprises with a defensible governance foundation while the vendor and regulatory accountability structures described above continue to develop.

References

[1] Jones Walker LLP. "AI Vendor Liability Squeeze: Courts Expand Accountability While Contracts Shift Risk." Jones Walker AI Law Blog, 2026.

[2] AI CERTs. "Independent Action Risk: Addressing Liability Gaps in Agentic AI." AI CERTs News, 2026.

[3] Baker Donelson. "2026 AI Legal Forecast: From Innovation to Compliance." Baker Donelson, 2026.

[4] Clifford Chance. "Agentic AI: The Liability Gap Your Contracts May Not Cover." Clifford Chance Talking Tech, February 2026.

[5] Cloudflare. "Scaling MCP Adoption: Our Reference Architecture for Simpler, Safer and Cheaper Enterprise Deployments of MCP." Cloudflare Blog, 2026.

[6] U.S. Office of Management and Budget. "M-26-04: Increasing Public Trust in Artificial Intelligence Through Unbiased AI Principles." OMB, December 11, 2025.

[7] Ropes & Gray. "Newsom Signs Executive Order Establishing AI Vendor Certification and Procurement Framework." Ropes & Gray Insights, April 2026.

[8] arXiv. "Securing the Model Context Protocol (MCP): Risks, Controls, and Governance." arXiv:2511.20920, November 2025.

[9] Invariant Labs. "MCP Security Notification: Tool Poisoning Attacks." Invariant Labs Blog, 2025.

[10] MCP Manager. "MCP Supply Chain Security & Risks." MCP Manager, 2026.

[11] Kiteworks. "MCP 'By Design' Flaw: The AI Supply Chain Risk CISOs Can't Ignore." Kiteworks, April 2026.

[12] OECD. "OECD Due Diligence Guidance for Responsible AI." OECD Publishing, February 2026.

[13] CPO Magazine. "2026 AI Legal Forecast: From Innovation to Compliance." CPO Magazine, 2026.

[14] Cloud Security Alliance. "Agentic AI Threat Modeling Framework: MAESTRO." CSA Blog, February 2025.

[15] Cloud Security Alliance. "CSA STAR for AI." Cloud Security Alliance, 2026.

[16] Cloud Security Alliance. "AI Organizational Responsibilities: Governance, Risk Management, Compliance and Cultural Aspects." CSA, 2024.

[17] Cloud Security Alliance. "The Agentic Trust Framework: Zero Trust Governance for AI Agents." CSA Blog, February 2026.

[18] Elish, Madeleine Clare. "Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction." Engaging Science, Technology, and Society 5, 2019.