AI AGENTS, DATA HANDLING & CYBERSECURITY

Contractual Risk Allocation for Autonomous AI Systems in Enterprise Environments

A Guide for Legal and Technology Professionals

2026 Edition

About the Author

Colin S. Levy

Colin S. Levy is a corporate lawyer, author, educator, and recognized leader in legal technology with over a decade of experience bridging the gap between technology and law. He serves as General Counsel at Malbek and as Adjunct Professor of Law at Albany Law School, where he teaches on the intersection of legal practice and technology innovation.

Colin is the author of *The Legal Tech Ecosystem*, co-author of *CLM for Dummies*, and editor of the *Handbook of Legal Tech*. His insights on legal innovation appear regularly in *Today's General Counsel* and other leading publications. He serves as a judge for the American Legal Technology Awards and is a frequent speaker at legal technology conferences worldwide.

A graduate of Trinity College and Boston College Law School, Colin also holds a Certificate in Legal Innovation and Technology from Suffolk University Law School. His career has been defined by a commitment to educating legal professionals on leveraging technology to improve the delivery of legal services and empowering the next generation of lawyers to embrace innovation.

Table of Contents

Executive Summary

Enterprise organizations are deploying AI agents under contracts written for passive, predictable software. This guide examines why that approach creates unacceptable risk and what legal and technology professionals should do about it.

The Core Problem

AI agents are not ordinary software. They make autonomous decisions about which data to access, which tools to invoke, and which actions to take. They chain operations across multiple enterprise systems in sequences no developer anticipated. They maintain persistent memory across sessions. Standard SaaS contracts, which assume the customer controls what the software does, are structurally incapable of governing these systems.

Key Security Risks

AI agents face a distinctive threat landscape. Prompt injection, ranked as the number one risk by OWASP, exploits the fact that agents cannot reliably distinguish legitimate instructions from malicious content in the data they process. All known defenses have been bypassed by adaptive attacks. Beyond prompt injection, agents are vulnerable to memory poisoning (where compromised memory persists across sessions), tool poisoning through MCP supply chains, semantic privilege escalation (where agents act within their technical permissions but outside their intended purpose), and data exfiltration through chained tool use.

Where Contracts Fail

Standard contractual protections break down across four dimensions. Fee-based liability caps are disproportionate to the potential harm agents can cause. Indemnification provisions do not cover agent-initiated IP infringement or autonomous actions. Data Processing Agreements that assume predetermined data flows cannot govern agents that decide at runtime which data to access. And major insurers are filing broad AI exclusions across D&O;, E&O;, and cyber policies, leaving organizations without traditional risk transfer mechanisms.

The Regulatory Landscape

Multiple regulatory frameworks now apply to AI agent deployments, including the EU AI Act (full applicability for high-risk systems in August 2026), the NIST AI Risk Management Framework, the OWASP Top 10 for LLM Applications and its companion framework for agentic AI, and state-level legislation including the Colorado AI Act and Illinois and Maine disclosure requirements. These frameworks create obligations that must be reflected in vendor contracts.

What to Do About It

Part V of this guide provides specific contractual provisions for agent-aware agreements. Key recommendations include: replacing "use of the service" language with enumerated delegation of

authority provisions; establishing escalation thresholds that require human approval for consequential actions; implementing purpose-locked data access with runtime enforcement; adopting fault-based liability allocation that distinguishes design failures from configuration errors and external attacks; mandating documented and tested kill switch mechanisms; and requiring AI-specific insurance coverage or alternative risk transfer arrangements.

Part One

The Problem

Why treating AI agents like ordinary software creates unacceptable risk

Why AI Agents Are Not Ordinary Software

Enterprise software contracts have evolved over decades to allocate risk between vendors and customers. Limitation of liability clauses, indemnification provisions, data processing agreements, and insurance requirements all assume a fundamental characteristic of the software they govern: predictability. Traditional software executes predefined commands in predetermined sequences, accesses specific datasets through defined interfaces, and produces outputs that are deterministic or at least bounded.[2]

AI agents violate every one of these assumptions. An AI agent is not a tool that waits for instructions. It is an autonomous system that reasons about goals, decides which tools to invoke, determines what data to access based on runtime analysis, and takes actions across enterprise systems with minimal human oversight. The distinction is not incremental. It is architectural, and it demands a fundamentally different approach to contractual risk allocation.[3]

Autonomous Decision-Making

Traditional software follows execution paths defined at development time. When a user clicks a button, the software performs the action mapped to that button. AI agents operate differently. Given a goal such as "prepare a summary of all customer complaints from the last quarter," an agent independently decides which databases to query, which documents to retrieve, which APIs to call, and how to synthesize the results. These decisions are made at runtime, not at design time.[2]

Dynamic Data Access Patterns

In traditional software, data flows are predetermined and documented. A CRM system accesses the customer database; a billing system accesses the financial database. Data processing agreements can specify exactly which datasets will be processed and for what purposes. AI agents break this model entirely. An agent tasked with resolving a customer issue might access the CRM, then the billing system, then internal knowledge bases, then external APIs, all based on its own assessment of what information is needed to complete the task.[25]

Chained Tool Use and Multi-System Access

Agents integrate external tools, APIs, databases, and services, inheriting all associated privileges. When an agent has access to an email API, a file system, and a database, it can chain these together in sequences never anticipated by the developers. A compromised or manipulated agent with access to multiple systems can execute attack chains that cross security boundaries in ways traditional malware cannot replicate.[2]

Persistent Memory Across Sessions

Unlike stateless traditional software, many AI agents maintain memory across sessions and conversations. This persistence enables continuity and personalization, but it also creates attack

vectors that do not exist in traditional systems. Once an agent's memory is compromised, the corruption persists until discovered and remediated.[7][8]

Key Risks

Unpredictable Scope: Agents determine their own scope of action at runtime. No contract provision can enumerate all possible actions an autonomous agent might take.

Cascading Consequences: A single agent decision can trigger chains of actions across multiple systems, each compounding the impact of errors or manipulation.

Attribution Difficulty: When an agent takes an unauthorized action, determining whether the cause was a prompt injection attack, a training data issue, a configuration error, or an inherent model limitation is technically and legally complex.

The Contractual Gap

Standard enterprise software agreements allocate risk around "use of the service," which assumes the customer controls what the software does. The customer initiates actions, the software executes them, and liability flows accordingly. With AI agents, many consequential actions are initiated by the agent itself, not by the customer. The fundamental question of who authorized a particular action becomes difficult to answer.[3][24]

As Professors Ian Ayres and Jack Balkin argued in the University of Chicago Law Review, the law of AI is the law of "risky agents without intentions." AI agents create risks of harm but lack the intentions that many legal frameworks use to assign liability. The legal system must therefore focus on the deployer's duty of care in configuring, monitoring, and constraining the agent, rather than on the agent's own "decisions."[23]

Under the Uniform Electronic Transactions Act (UETA) and E-SIGN, AI agents that initiate transactions and respond to electronic records may qualify as "electronic agents" capable of forming enforceable contracts. The general principle under UETA is that the party deploying an electronic agent bears responsibility for the results it produces, because the agent lacks independent volition of its own.[24]

This legal backdrop exposes a critical gap: organizations are deploying autonomous systems under contracts designed for passive tools. The result is that liability for agent-initiated harm falls into contractual gray areas where neither party has clearly accepted responsibility, while insurance markets are actively retreating from coverage.

Contractual Considerations

Scope Ambiguity: Standard 'use of the service' language does not capture autonomous agent actions taken without explicit user authorization.

Liability Mismatch: Fee-based liability caps designed for predictable SaaS services are inadequate for agents capable of causing cascading enterprise-wide harm.

Data Processing Gaps: DPAs that assume predetermined data flows cannot govern agents that determine data access dynamically at runtime.

Insurance Retreat: Major insurers are adding broad AI exclusions to D&O;, E&O;, and cyber policies, leaving organizations without traditional risk transfer mechanisms.

Part Two

Cybersecurity Risks

Threats unique to autonomous AI agents

Prompt Injection and Indirect Prompt Injection

Prompt injection is the most critical vulnerability facing AI agents, ranked as the number one risk in OWASP's 2025 Top 10 for LLM Applications.[1] Unlike traditional software exploits that target code vulnerabilities, prompt injection targets the reasoning process of the AI system itself. The attack exploits the fact that AI agents cannot reliably distinguish between legitimate instructions and malicious content embedded in the data they process.

Direct prompt injection involves an attacker submitting malicious instructions through the visible input interface. Indirect prompt injection is far more dangerous for agents: malicious instructions are embedded in the data the agent processes, including web pages, documents, emails, database records, MCP tool descriptions, or memory entries.[4] When an agent retrieves and processes this data, it may interpret the embedded instructions as legitimate directives and act on them.

A 2024 evaluation of 30 different LLM agents found that agents are vulnerable to indirect prompt injection attacks at significant rates, with ReAct-prompted GPT-4 vulnerable 24% of the time. Enhanced attack techniques increased success rates substantially.[29] A 2025 study published at NAACL evaluated eight different defense mechanisms against adaptive attacks and bypassed all of them, consistently achieving attack success rates above 50%.[4]

Key Risks

No Reliable Defense: Current defenses against indirect prompt injection remain unreliable. All known mitigation strategies have been bypassed by adaptive attacks.

Contractual Blind Spot: Standard contracts do not address liability for actions an agent takes as a result of processing maliciously crafted content encountered during normal operation.

Data Exfiltration Through Agent Tool Use

AI agents with access to communication tools, file systems, and APIs can be manipulated into exfiltrating sensitive data. Because agents combine data retrieval with the ability to take external actions such as sending messages, writing files, or calling APIs, they can serve as unwitting exfiltration channels.

In August 2024, researchers discovered that Slack AI could be exploited through a combination of RAG poisoning and social engineering. By crafting malicious content with hidden Markdown links in public channels, attackers tricked Slack AI into summarizing sensitive conversations from private channels and transmitting those summaries to attacker-controlled destinations. This attack exploited the agent's dynamic data access pattern: it accessed whatever channels were relevant to its summarization task.[2]

Palo Alto Networks Unit 42 has documented scenarios in which financial services agents were manipulated into exporting large volumes of customer records by crafting requests that appeared to be legitimate business tasks. Because the agent determines what data to retrieve based on its own reasoning, a well-crafted prompt can cause it to treat a mass export as routine.[2]

Contractual Considerations

Data Boundary Definitions: Contracts should specify explicit data boundaries, defining which datasets agents can and cannot access, with technical enforcement rather than policy-only controls.

Exfiltration Liability: Responsibility for agent-mediated data exfiltration should be allocated based on whether the vulnerability resulted from agent design, deployment configuration, or external attack.

Privilege Escalation and Semantic Privilege Escalation

When AI agents gain access to APIs, databases, and external services, they inherit all associated privileges. Traditional privilege escalation occurs when an agent is manipulated into accessing resources beyond its authorized scope.[2] Unit 42 researchers documented successful attacks against popular agent frameworks where agents were manipulated to perform SQL injection, steal service account tokens from cloud metadata endpoints, and exfiltrate credentials from mounted volumes.

Semantic privilege escalation is a newer and subtler threat. It occurs when an agent is technically authorized to access a resource but uses that access in unintended ways. For example, an agent authorized to read an email inbox for scheduling purposes could be manipulated into reading confidential communications and forwarding their contents externally.[6] The agent has not exceeded its technical permissions, but it has exceeded the intended scope of its access, creating a category of harm that traditional access controls cannot prevent.

In multi-agent architectures, risk compounds. An orchestration agent holding API keys for downstream agents represents a single point of compromise. If the orchestrator is manipulated, the attacker gains access to all downstream systems simultaneously.[2]

Key Risks

Permission Inheritance: Agents inherit all permissions of the accounts they authenticate through. Standard role-based access controls do not account for an agent's ability to chain permissions across systems.

Semantic vs. Technical Authorization: An agent can act within its technical permissions while violating the intended purpose of those permissions, a gap that access control lists and firewalls cannot detect.

Memory Poisoning and Persistent Compromise

AI agents with persistent memory introduce a category of vulnerability with no analog in traditional software. Memory poisoning involves injecting malicious instructions into an agent's long-term memory, causing the agent to behave differently in all future interactions without any visible indication of compromise.[7]

At NeurIPS 2024, researchers from the University of Chicago, UIUC, UW-Madison, and UC Berkeley presented AgentPoison, the first backdoor attack framework targeting RAG-based LLM agents by poisoning their knowledge bases.[7] A separate 2025 study demonstrated the MINJA attack, showing that agents with persistent memory are vulnerable to progressive injection techniques achieving injection success rates above 98%.[8]

In September 2024, researchers demonstrated that ChatGPT's memory could be exploited to create persistent "spAIware." Hidden instructions embedded in websites were injected into a user's ChatGPT memory through seemingly innocent requests like "summarize this webpage." Once injected, the instructions survived across chat sessions and device changes because memories are stored server-side. In January 2026, Radware researchers extended this work with "ZombieAgent," demonstrating that ChatGPT's connector and memory features can be combined to make indirect prompt injection attacks persistent and cross-session.

Contractual Considerations

Memory Audit Rights: Contracts should include rights to audit agent memory contents, with required logging of all memory modifications and mechanisms for purging compromised memories.

Persistence Liability: Liability allocation should address the extended duration of memory-based compromises, which may persist for weeks or months before detection.

Supply Chain Risks: MCP Servers and Tool Ecosystems

The Model Context Protocol (MCP), launched by Anthropic in November 2024, standardizes how AI agents connect to external data sources and tools. While MCP has driven rapid ecosystem growth, it has also introduced significant supply chain vulnerabilities.[10]

Tool poisoning occurs when malicious instructions are embedded within MCP tool descriptions. These instructions are invisible to users but visible to AI models. An attacker can publish a seemingly benign tool with a description containing hidden directives, such as instructions to read SSH keys and pass their contents as parameters.[5] When an agent loads and parses the tool description, it treats these hidden instructions as legitimate operational guidance.

Security researchers analyzing publicly available MCP server implementations in 2025 found that 43% of tested implementations contained command injection flaws, 30% were vulnerable to server-side request forgery, and 22% leaked files outside intended directories.[10] CVE-2025-6514 affected mcp-remote, a trusted OAuth proxy used by over 437,000 developers. A poisoned update compromised all users in what security researchers termed a "rug pull."[9]

The contractual challenge is that responsibility flows through multiple parties. In a typical agent deployment, the chain runs from user to agent provider to tool/plugin provider to data source. Standard contracts do not address how liability cascades when a compromise originates in a third-party tool that the agent provider integrated and the user never directly authorized.

Key Risks

Tool Verification Gap: No standardized mechanism exists for verifying the safety of MCP tool descriptions before an agent processes them.

Multi-Party Liability: Supply chain compromises cross contractual boundaries, creating attribution and liability gaps between agent providers, tool providers, and users.

Excessive Agency

Excessive agency, ranked as LLM06 in the OWASP 2025 Top 10, occurs when an AI agent is granted more power, permissions, or system-level capabilities than its task actually requires. OWASP identifies three root causes: excessive functionality, excessive permissions, and excessive autonomy.[11]

Excessive functionality means an agent has access to tools or extensions beyond what its intended operation requires. For example, an agent designed to search and summarize documents should not also have the ability to send emails or modify files. Excessive permissions occur when an agent operates with broader access rights than necessary, such as a read-only task executing under an account with write and delete privileges. Excessive autonomy means an agent takes consequential actions without sufficient human oversight or approval checkpoints.

The December 2025 OWASP Top 10 for Agentic Applications further developed this framework, introducing the principle of "least agency": only grant agents the minimum autonomy required to perform safe, bounded tasks.[26] This principle should be a contractual requirement, not merely a design recommendation.

Implementation Guidance

Least Privilege Enforcement: Contracts should require that agent permissions are scoped to the minimum necessary for each specific task, with technical enforcement mechanisms rather than policy-only controls.

Function Restriction: Agent access to tools and extensions should be explicitly enumerated in contracts, with any expansion requiring written authorization.

Autonomy Boundaries: Contracts should define clear thresholds where agent actions require human approval, with escalation procedures for ambiguous cases.

Part Three

Where Contracts Fail

How standard terms create unacceptable gaps

Limitation of Liability Clauses

Standard SaaS agreements cap vendor liability at the fees paid during the preceding 12 months. This approach reflects an assumption that the potential harm from software failure is roughly proportional to its cost. For a subscription costing several thousand dollars per year, the exposure is bounded and manageable.

AI agents shatter this proportionality. An agent with access to enterprise systems can cause harm that exceeds its subscription cost by orders of magnitude. A misconfigured or compromised agent could exfiltrate an entire customer database, execute unauthorized financial transactions, or trigger regulatory violations across multiple jurisdictions, all within minutes. The standard "low cap plus broad disclaimer" approach is under significant pressure.[3]

Data from TermScout shows that 88% of AI vendors impose liability caps, yet only 38% cap customer liability, compared to 44% in broader SaaS agreements. This asymmetry means vendors have limited their own exposure while leaving customers with disproportionate risk.[18]

The emerging legislative response reflects recognition of this problem. The AI LEAD Act prohibits developers from imposing contract provisions that unreasonably limit liability on deployers, and declares such clauses unenforceable.

Contractual Considerations

Tiered Liability Caps: Contracts should establish differentiated caps based on risk category. Agent-initiated data breaches, regulatory violations, and autonomous actions should carry higher caps than standard service failures.

Carve-Outs for Agent Actions: Specific carve-outs should exclude agent-initiated harm from general liability caps, particularly for data breaches and regulatory non-compliance resulting from autonomous agent decisions.

Uncapped Categories: Certain categories of agent-caused harm, including willful misconduct facilitated by agent misconfiguration and failure to implement required safety controls, should remain uncapped.

Indemnification Gaps

Traditional indemnification provisions address three categories: intellectual property infringement, data breaches, and negligence. Each category assumes identifiable human decisions and controllable outcomes. AI agents introduce scenarios that fall outside all three.[19]

When an AI agent generates output that infringes a third party's intellectual property, the indemnification question is immediate. Yet most AI vendors either exclude or severely limit IP infringement indemnification for AI-generated outputs.[31] Many vendors explicitly exclude training data and model outputs from their indemnification obligations, leaving customers exposed to infringement claims for content they neither created nor controlled.

The challenge deepens with agent-initiated actions. If an agent autonomously accesses a dataset, generates a document incorporating protected content, and distributes it to a client, the indemnification chain involves the model provider, the agent platform, the tool providers that supplied the data, and the deploying organization. Standard contracts address none of these multi-party scenarios.

AI providers commonly block the flow of warranties and indemnities to end clients. Even broad indemnification coupled with a relatively low liability cap renders protection quite limited or effectively illusory.[19]

Contractual Considerations

Output Indemnification: Contracts should clearly allocate indemnification for IP infringement in agent-generated outputs, specifying responsibility across the entire chain from model provider to deployer.

Action Indemnification: Indemnification for harm caused by autonomous agent actions should be addressed separately from traditional product liability, with clear allocation based on the source of the failure.

Flow-Through Protections: Organizations should negotiate flow-through indemnification rights that extend protections from upstream model and tool providers to the deploying organization.

Data Processing Agreements

Standard Data Processing Agreements under GDPR and similar frameworks assume deterministic data handling. They specify the exact datasets to be processed, the specific purposes for processing, and the predetermined data flows between systems. An AI agent operating autonomously breaks every one of these assumptions.[25]

The GDPR's principles of purpose limitation, data minimization, transparency, storage limitation, and accountability remain fully applicable to AI agents. However, as the IAPP has analyzed, "the operating model around those principles, including stable data flows, predictable toolchains, and human approvals, is what fails."[25] When an agent determines at runtime that it needs to access additional datasets to complete a task, the static purpose specification in a traditional DPA cannot govern that access.

The European Data Protection Board issued Opinion 28/2024 in December 2024 on data protection aspects of AI models, emphasizing that Articles 13 and 14 require clear information about data processing. With agentic AI, processing operations may change dynamically as agents adapt.[30] The IAPP recommends shifting from static documents to runtime enforcement mechanisms, including purpose locks that constrain agent data access to approved categories, execution traces that log all agent data interactions, and controller-processor cartography that maps data flows dynamically.[25]

Implementation Guidance

Purpose-Locked Access: DPAs should define categories of permitted data access rather than specific datasets, with runtime enforcement mechanisms that restrict agent behavior to approved purposes.

Execution Tracing: Contracts should require comprehensive logging of all agent data access events, with audit rights that enable verification of compliance with stated processing purposes.

Dynamic Consent Mechanisms: For agents exposed to third-party plugins, contracts should require trace-level logging, deletion APIs, and Data Protection Impact Assessments.

The Insurance Coverage Crisis

The insurance industry's response to AI agent risk represents a significant departure from traditional technology risk transfer. Rather than adapting coverage to address new risks, major carriers are withdrawing from AI liability entirely.[21]

Major insurers including AIG, Great American, and WR Berkley have filed with state regulators for broad generative AI exclusions or moved to exclude AI-related Errors and Omissions claims. WR Berkley's filed exclusion would eliminate coverage for claims arising out of AI use, deployment, or development, AI-generated content, failure to detect AI-produced materials, inadequate AI governance, chatbot communications, and regulatory actions related to AI oversight.[21] Such exclusions can apply across D&O;, E&O;, and fiduciary liability policies, though implementation status varies by carrier.

Insurers argue that AI risk is not actuarially mature and that a single faulty model update or misconfigured agent can produce simultaneous, widespread losses. This pattern resembles catastrophe exposure rather than traditional liability, and the industry is unwilling to absorb unbounded exposure while pricing models remain undeveloped.[22]

Specialty insurers are beginning to fill the gap. Relm Insurance has introduced PONTAAI, an excess difference-in-conditions wrap policy for AI liability. Armilla Insurance Services offers AI liability coverage underwritten by Lloyd's of London syndicates, addressing AI-specific perils such as hallucinations, model drift, and mechanical failures.[33] However, these specialty products remain limited in availability and often expensive relative to the coverage they provide.

Key Risks

Coverage Gaps: Organizations deploying AI agents may discover that their existing cyber, D&O;, and E&O; policies exclude AI-related claims entirely, leaving them without risk transfer mechanisms for the most consequential category of exposure.

Contractual Compensation: In the absence of adequate insurance, contracts must bear a greater share of risk allocation. Indemnification provisions and liability caps become the primary risk transfer mechanism.

Part Four

Regulatory Landscape

Standards and requirements shaping AI agent governance

EU AI Act

The EU AI Act entered into force on August 1, 2024, with full applicability for high-risk AI systems on August 2, 2026.[12] The Act establishes a risk-based classification system with escalating obligations. AI systems used in biometrics, critical infrastructure, education, employment, essential services, law enforcement, migration, and justice are classified as high-risk under Annex III.[12]

Under Articles 8 through 16, providers of high-risk AI systems must conduct data governance to ensure training datasets are relevant and representative, draw up technical documentation demonstrating compliance, provide instructions for use to downstream deployers, design systems that allow human oversight, and automatically record events relevant for identifying risks throughout the system lifecycle.[13]

For AI agents specifically, classification may be fluid. Because agentic systems can self-update and acquire new capabilities, their risk levels may evolve after deployment, requiring continuous monitoring and change control procedures.[35] This dynamic risk profile means contractual obligations tied to a specific risk classification at deployment may become inadequate as the agent's capabilities change over time.

Implementation Guidance

Contractual Compliance Allocation: Contracts should specify which party bears responsibility for each EU AI Act obligation, including documentation, monitoring, and human oversight requirements.

Dynamic Classification: Provisions should address how contractual obligations adjust when an agent's risk classification changes due to capability evolution.

NIST AI Risk Management Framework

NIST released the AI Risk Management Framework (AI RMF 1.0) in January 2023 and supplemented it with the Generative AI Profile (NIST AI 600-1) in July 2024. The framework organizes AI risk management around four core functions: Govern, Map, Measure, and Manage.[14][15]

The Govern function establishes governance structures for responsible AI development and deployment. The Map function identifies and documents the AI system's lifecycle characteristics and risk context. The Measure function evaluates AI system performance and risk through quantitative and qualitative methods. The Manage function implements risk mitigation strategies identified through measurement.

The Generative AI Profile extends the base framework to address risks specific to large language models and generative systems, including confabulation, data privacy, environmental costs, and information security. For organizations deploying AI agents, the NIST framework provides a structured approach to risk identification that can inform contractual requirements and governance obligations.

OWASP Top 10 for LLM Applications

The OWASP Top 10 for LLM Applications, developed with over 600 contributing experts from more than 18 countries, provides the most widely referenced security framework for LLM systems.[1] The 2025 edition identifies the following top risks, in order: prompt injection, sensitive information disclosure, supply chain vulnerabilities, data and model poisoning, improper output handling, excessive agency, system prompt leakage, vector and embedding weaknesses, misinformation, and unbounded consumption.

In December 2025, OWASP released a companion framework specifically addressing agentic AI: the Top 10 for Agentic Applications.[26] This framework identifies agent goal hijacking, insecure inter-agent communication, cascading failures, and human-agent trust exploitation as critical risks specific to autonomous AI systems.

State-Level AI Legislation

The Colorado Artificial Intelligence Act (SB24-205), with an effective date postponed to June 30, 2026 by SB 25B-004, represents the most comprehensive U.S. state framework for AI governance.[16] The Act applies to developers and deployers of high-risk AI systems in employment, housing, financial services, insurance, and healthcare. Developers must make documentation available describing reasonably foreseeable uses and harmful uses, while deployers must implement risk management policies aligned with the NIST AI RMF and ISO 42001.

The Colorado Act requires annual impact assessments and assessments within 90 days of substantial modifications. Assessments must include the system's purpose, intended use cases, known risks, limitations, data inputs and outputs, performance metrics, and transparency measures. The Colorado Attorney General has exclusive enforcement authority.

Other states are following. Illinois mandates disclosure when AI influences employment decisions, effective January 1, 2026. Maine requires notification when consumers interact with AI agents rather than humans. These state-level requirements create contractual obligations that must be reflected in vendor agreements, as deployers need assurance that their AI agent vendors can support compliance with applicable state requirements.

GDPR and Autonomous Data Processing

The European Data Protection Board issued Opinion 28/2024 in December 2024, addressing AI and GDPR compliance and emphasizing challenges of transparency and purpose limitation for dynamic AI systems.[30] For AI agents that determine their own data access patterns at runtime, the traditional GDPR compliance model of documenting data flows in advance and obtaining purpose-specific consent is structurally inadequate.

The IAPP has identified four runtime controls essential for GDPR compliance in agentic systems:[25] purpose locks that constrain agent data access to approved categories, execution traces that provide complete audit trails of agent data interactions, memory controls that govern how agents retain and use personal data across sessions, and controller-processor cartography that maps the data protection roles across complex agent architectures.

ISO/IEC 42001:2023, the first international AI management system standard,[17] provides a complementary governance framework with 38 distinct controls including data quality requirements, human oversight mechanisms, and accuracy standards. Organizations deploying AI agents should consider ISO 42001 certification as both a governance foundation and a contractual compliance benchmark.

Part Five

Agent-Aware Contracts

Provisions that address the reality of autonomous AI systems

Redefining Scope and Delegation of Authority

The most fundamental change required in AI agent contracts is a shift from defining what the software does to defining what the agent is authorized to do. As Mayer Brown's 2026 analysis identifies, the contracting model must shift from SaaS with limited performance guarantees to a service-oriented model with explicit delegation of authority and policy guardrails.[3]

Delegation of authority provisions should outline precisely what the agent can and cannot do, specified at a granular level. Rather than broad language permitting the agent to "process data as necessary to provide the service," contracts should enumerate permitted action categories, data access boundaries, and decision-making authority limits.

Policy guardrails define how the agent operates within its delegated authority. These provisions specify mandatory escalation triggers for human-in-the-loop approval, operational constraints such as transaction limits and data sensitivity thresholds, and monitoring requirements that enable real-time oversight. The critical contractual question becomes: "What is the exact threshold where the AI agent must stop and ask a human for approval?"

Performance standards should shift from availability-based SLAs measuring uptime percentage to outcome-based metrics measuring the quality, accuracy, and safety of agent actions. Service credits should be triggered by performance failures, not just downtime.

Contractual Considerations

Enumerated Authorities: Contracts should list the specific categories of actions the agent is authorized to take, with any expansion requiring written amendment.

Escalation Thresholds: Define clear, measurable thresholds that trigger mandatory human review before the agent can proceed, including financial limits, data sensitivity levels, and irreversibility criteria.

Outcome-Based SLAs: Replace availability metrics with accuracy, safety, and quality metrics that reflect the actual performance characteristics that matter for autonomous agents.

Dynamic Data Governance Provisions

Traditional data governance provisions specify which data the service will process. For AI agents, contracts must instead establish boundaries and enforcement mechanisms that govern dynamic data access decisions the agent makes at runtime.

Data boundary specifications should define which systems the agent can access, which data classifications it can process, and which sensitivity levels require additional authorization. Technical enforcement should complement contractual provisions: rather than relying on the agent to respect data boundaries, organizations should implement architectural controls that prevent unauthorized access.

Permission inheritance rules must address whether an agent inherits all permissions of the user account it authenticates through or operates under a restricted permission set. In most cases, agents should operate under purpose-specific credentials with narrower permissions than any individual user, reflecting the principle of least privilege applied to autonomous systems.

Contracts should specify data ownership unambiguously. Under the BPO-style approach emerging for agent contracts, the customer owns all data submitted to or obtained by the agent, and all outputs created by the agent in performance of the service.[3] This should extend to prohibitions on using customer data to train models without explicit, separate consent.

Implementation Guidance

Access Classification Matrix: Create a matrix mapping data sensitivity levels to agent permission levels, with higher sensitivity requiring additional authorization or human-in-the-loop approval.

Credential Scoping: Deploy purpose-specific service accounts for agent access with permissions narrower than individual user accounts, enforced at the infrastructure level.

Training Data Restrictions: Contractually prohibit use of customer data for model training, fine-tuning, or capability improvement without separate, explicit written consent.

Cascading Liability Frameworks

Agent-initiated harm rarely results from a single action. Instead, a chain of agent decisions produces cascading consequences that cross system boundaries and compound impact at each step. Contracts must address liability for these multi-step failure modes.

Liability allocation should distinguish between several categories of agent failure. Design failures, where the agent's architecture creates foreseeable risks, should fall primarily on the agent provider. Configuration failures, where the deployer improperly set up permissions or guardrails, should fall primarily on the deployer. External attacks, where a prompt injection or tool poisoning exploit manipulates the agent, require a more nuanced allocation based on whether the vulnerability was in the agent's design, the deployment configuration, or a third-party tool.

Gartner predicts that over 40% of agentic AI projects will be canceled by the end of 2027 due to escalating costs, unclear business value, or inadequate risk controls.[27] This projection underscores the importance of establishing clear liability frameworks before deployment rather than attempting to resolve responsibility after incidents occur.

Contractual Considerations

Fault-Based Allocation: Liability should be allocated based on the source of the failure: design defect, configuration error, external attack, or third-party component failure.

Cascading Impact Provisions: Contracts should address how liability scales when a single agent error produces cascading consequences across multiple systems and business functions.

Shared Responsibility: Multi-party deployments should include explicit responsibility matrices mapping each party's obligations for prevention, detection, and remediation.

Kill Switches and Human Oversight Requirements

Every AI agent deployment should include a documented, tested mechanism for immediately halting agent operations. This is not merely a technical recommendation. The EU AI Act requires that high-risk AI systems be designed to allow deployers to implement human oversight, and multiple regulatory frameworks mandate the ability to interrupt autonomous operations.[13]

Contractual kill switch provisions should specify several elements: the technical mechanism for halting agent operations, the escalation procedures defining who is authorized to invoke the kill switch and under what circumstances, the handling of in-flight operations when an agent is stopped mid-task, and the testing and verification requirements ensuring the kill switch functions reliably.

Human oversight requirements should distinguish between human-in-the-loop, where human approval is required before the agent acts, and human-on-the-loop, where humans monitor agent actions with the ability to intervene. The appropriate level of oversight should be tied to the consequence severity of the agent's actions. Irreversible actions, actions affecting sensitive data, and actions with external impact should require human-in-the-loop approval.

Industry practitioners have identified a significant governance-containment gap: many organizations that have implemented monitoring and human oversight capabilities for AI agents have not yet deployed corresponding containment controls, including kill switch mechanisms. This gap represents a critical vulnerability for organizations that can detect agent misbehavior but cannot reliably stop it.

Contractual Considerations

Kill Switch Requirements: Contracts should mandate documented, tested kill switch mechanisms with defined invocation authority, response times, and in-flight operation handling procedures.

Oversight Level Specification: Define which agent actions require human-in-the-loop approval versus human-on-the-loop monitoring, tied to consequence severity classifications.

Testing Requirements: Kill switch and human oversight mechanisms should be tested regularly, with testing frequency and documentation requirements specified contractually.

Incident Response for Agent-Initiated Breaches

Agent-initiated breaches differ from traditional cybersecurity incidents in several critical ways. The agent may not recognize its own compromise. The breach may unfold over multiple sessions as poisoned memory persists. The scope of affected data may be difficult to determine because the agent's data access was dynamic rather than predetermined.

IBM's 2025 report found that 13% of organizations reported breaches of AI models or applications, with 97% of those breached reporting that they lacked proper AI access controls. 63% of breached organizations either did not have an AI governance policy or were still developing one.[20] Additionally, one in five organizations reported a breach due to shadow AI, with organizations experiencing high levels of shadow AI observing breach costs $670,000 higher than those with low or no shadow AI.

Gartner predicts that AI agents will reduce the time it takes to exploit account exposures by 50% by 2027,[28] compressing the window organizations have to detect and respond to compromises. Incident response procedures designed for human-speed attacks may be inadequate for agent-speed exploitation.

Implementation Guidance

Agent-Specific IR Plans: Incident response plans should include agent-specific runbooks covering memory compromise detection, tool poisoning identification, and dynamic data access scope assessment.

Behavioral Logging: Contracts should require comprehensive logging of all agent actions, tool invocations, data accesses, and decision points, with retention periods sufficient for forensic analysis.

Containment Procedures: IR plans should define specific containment steps for agent incidents, including memory purging, credential rotation for all agent-accessible systems, and tool access revocation.

Notification Timelines: Breach notification obligations should account for the extended discovery timelines associated with persistent memory compromises and gradual data exfiltration.

Insurance and Risk Transfer

With traditional insurance markets retreating from AI liability, organizations must develop alternative risk transfer strategies. Contractual provisions must bear more of the risk allocation burden that insurance traditionally absorbed.

Organizations should evaluate specialty AI insurance products where available. Products such as Relm's PONTAAI policy and Armilla's Lloyd's-backed coverage address AI-specific perils that traditional policies exclude. However, these products remain limited in availability and coverage scope.[33]

Vendor contracts should require the agent provider to maintain insurance covering AI-related liabilities, with minimum coverage amounts that reflect actual exposure rather than nominal amounts. Contracts should specify that insurance requirements include coverage for agent-initiated data breaches, autonomous decision errors, and third-party claims arising from agent actions.

Risk retention strategies should include escrow arrangements, performance bonds, or retained earnings reserves sized to the estimated exposure from agent deployment. Organizations should also consider whether the total cost of risk, including insurance premiums, contractual risk retention, and residual uninsured exposure, justifies the deployment of autonomous agents versus more controlled alternatives.

Contractual Considerations

Vendor Insurance Requirements: Require agent providers to maintain AI-specific liability insurance with minimum coverage amounts reflecting actual deployment risk, not nominal SaaS-level amounts.

Risk Transfer Mechanisms: Where insurance is unavailable or insufficient, contracts should incorporate alternative mechanisms such as performance bonds, escrow arrangements, or enhanced indemnification.

Coverage Verification: Contracts should include rights to verify vendor insurance coverage annually and require notice of any material changes to coverage, including AI exclusion endorsements.

Glossary of Key Terms

AI Agent

An autonomous software system that reasons about goals, selects tools and data sources at runtime, and takes actions across enterprise systems with minimal human direction. Distinguished from traditional software by its capacity for independent decision-making.

Chained Tool Use

The ability of an AI agent to invoke multiple tools, APIs, or services in sequence, where each action builds on the results of previous actions. Creates attack surfaces that cross security boundaries between systems.

D&O / E&O Insurance

Directors and Officers (D&O) liability insurance and Errors and Omissions (E&O) professional liability insurance. Traditional policies that major insurers are now filing AI-related exclusions against.

Data Processing Agreement (DPA)

A contract required under GDPR and similar frameworks governing how a service provider processes personal data on behalf of a controller. Traditional DPAs assume predetermined, static data flows.

Electronic Agent

Under the Uniform Electronic Transactions Act (UETA), a computer program or electronic means used to initiate an action or respond to electronic records without human review. AI agents may qualify as electronic agents capable of forming enforceable contracts.

Excessive Agency

OWASP LLM06:2025. Occurs when an AI agent is granted more functionality, permissions, or autonomy than its task requires. Encompasses excessive functionality, excessive permissions, and excessive autonomy.

Human-in-the-Loop / Human-on-the-Loop

Two levels of human oversight. Human-in-the-loop requires human approval before the agent acts. Human-on-the-loop allows the agent to act while humans monitor with the ability to intervene. The appropriate level depends on the consequence severity of the action.

Indirect Prompt Injection

An attack in which malicious instructions are embedded in data the agent processes (web pages, documents, emails, database records) rather than submitted directly by a user. When the agent retrieves this data, it may follow the embedded instructions as if they were legitimate directives.

Kill Switch

A documented, tested mechanism for immediately halting AI agent operations. Required by the EU AI Act for high-risk systems and recommended as a contractual requirement for all agent deployments.

Least Agency

A principle introduced by the OWASP Top 10 for Agentic Applications (2025). States that AI agents should be granted the minimum autonomy, tool access, and credential scope required to perform their intended task. The agentic equivalent of the principle of least privilege.

MCP (Model Context Protocol)

An open protocol launched by Anthropic in November 2024 that standardizes how AI agents connect to external data sources and tools. MCP servers provide tools and data; MCP clients (agents) consume them. The ecosystem introduces supply chain risks including tool poisoning.

Memory Poisoning

An attack that injects malicious instructions into an AI agent's persistent memory, causing altered behavior in all future interactions. Unlike traditional exploits, the compromise persists across sessions until the memory is specifically purged.

Prompt Injection

The top-ranked risk in OWASP's 2025 Top 10 for LLM Applications. An attack that targets the reasoning process of an AI system by submitting malicious instructions that the model cannot distinguish from legitimate input. See also: Indirect Prompt Injection.

RAG (Retrieval-Augmented Generation)

A technique where an AI model retrieves relevant documents or data from external sources before generating a response. Creates attack surfaces when retrieved content contains malicious instructions (RAG poisoning).

Semantic Privilege Escalation

An attack in which an agent is technically authorized to access a resource but uses that access in unintended ways. For example, an agent authorized to read emails for scheduling purposes reads confidential communications instead. Traditional access controls cannot detect this category of misuse.

Tool Poisoning

An attack in which malicious instructions are embedded within MCP tool descriptions. These instructions are invisible to human users but visible to AI models. When an agent loads the tool, it treats the hidden instructions as legitimate operational guidance.

Endnotes

[1] OWASP Foundation, "OWASP Top 10 for LLM Applications 2025," 2025. Available at: https://genai.owasp.org/resource/owasp-top-10-for-llm-applications-2025/

[2] Palo Alto Networks Unit 42, "AI Agents Are Here. So Are the Threats," 2024. Available at: https://unit42.paloaltonetworks.com/agentic-ai-threats/

[3] Mayer Brown, "Contracting for Agentic AI Solutions: Shifting the Model from SaaS to Services," February 2026. Available at: https://www.mayerbrown.com/en/insights/publications/2026/02/contracting-for-agentic-ai-solutions-shifting-the-model-from-saas-to-services

[4] Qiusi Zhan et al., "Adaptive Attacks Break Defenses Against Indirect Prompt Injection Attacks on LLM Agents," Findings of NAACL 2025. Available at: https://aclanthology.org/2025.findings-naacl.395/

[5] Acuvity, "Tool Poisoning: Hidden Instructions in MCP Tool Descriptions," 2024. Available at: https://acuvity.ai/tool-poisoning-hidden-instructions-in-mcp-tool-descriptions/

[6] Acuvity, "Semantic Privilege Escalation: The Agent Security Threat Hiding in Plain Sight," 2024. Available at: https://acuvity.ai/semantic-privilege-escalation-the-agent-security-threat-hiding-in-plain-sight/

[7] NeurIPS 2024, "AGENTPOISON: Red-teaming LLM Agents via Poisoning Memory or Knowledge Bases," University of Chicago et al. Available at: https://proceedings.neurips.cc/paper_files/paper/2024/file/eb113910e9c3f6242541c1652e30dfd6-Paper-Conference.pdf

[8] Zhen Tan et al., "MINJA: Memory Injection Attacks on LLM Agents via Query-Only Interaction," 2025. Available at: https://arxiv.org/abs/2503.03704

[9] Docker, "MCP Horror Stories: The Supply Chain Attack," 2025. Available at: https://www.docker.com/blog/mcp-horror-stories-the-supply-chain-attack/

[10] Pillar Security, "The Security Risks of Model Context Protocol (MCP)," 2024. Available at: https://www.pillar.security/blog/the-security-risks-of-model-context-protocol-mcp

[11] OWASP Gen AI Security Project, "LLM06:2025 Excessive Agency," 2025. Available at: https://genai.owasp.org/llmrisk/llm062025-excessive-agency/

[12] European Union, "EU AI Act Article 6: Classification Rules for High-Risk AI Systems," 2024. Available at: https://artificialintelligenceact.eu/article/6/

[13] European Union, "EU AI Act Articles 8-16: Requirements and Obligations for High-Risk AI Systems," 2024. Available at: https://artificialintelligenceact.eu/article/16/

[14] NIST, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," January 2023. Available at: https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf

[15] NIST, "AI RMF Generative Artificial Intelligence Profile (NIST AI 600-1)," July 2024. Available at: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf

[16] Colorado General Assembly, "SB24-205: Consumer Protections for Artificial Intelligence," 2024; effective date postponed to June 30, 2026 by SB 25B-004 (signed August 28, 2025). Available at: https://leg.colorado.gov/bills/sb24-205

[17] ISO, "ISO/IEC 42001:2023 - Artificial Intelligence Management System," 2023. Available at: https://www.iso.org/standard/42001

[18] Scott & Scott LLP, "Limitations of Liability in Artificial Intelligence Contracts," 2024. Available at: https://scottandscottllp.com/limitations-of-liability-in-artificial-intelligence-contracts/

[19] Parsons Behle & Latimer, "Indemnification Clauses in Contracts Involving Artificial Intelligence," 2024. Available at: https://parsonsbehle.com/insights/indemnification-clauses-in-contracts-involving-artificial-intelligence-how-well-is-your-business-protected

[20] IBM, "IBM Report: 13% Of Organizations Reported Breaches Of AI Models Or Applications," July 2025. Available at: https://newsroom.ibm.com/2025-07-30-ibm-report-13-of-organizations-reported-breaches-of-ai-models-or-applications

[21] Hunton Andrews Kurth, "The Continued Proliferation of AI Exclusions," 2024. Available at: https://www.hunton.com/hunton-insurance-recovery-blog/the-continued-proliferation-of-ai-exclusions

[22] American Bar Association, "The Evolving Landscape of AI Insurance: Empirical Insights into Risks and Policy Gaps," 2025. Available at: https://www.americanbar.org/groups/tort_trial_insurance_practice/resources/brief/2025-fall/evolving-landscape-ai-insurance-empirical-insights-risks-policy-gaps/

[23] Ian Ayres & Jack M. Balkin, "The Law of AI is the Law of Risky Agents Without Intentions," University of Chicago Law Review Online, June 2024. Available at: https://lawreview.uchicago.edu/online-archive/law-ai-law-risky-agents-without-intentions

[24] Proskauer Rose, "Contract Law in the Age of Agentic AI: Who's Really Clicking Accept?" April 2025. Available at: https://newmedialaw.proskauer.com/2025/04/09/contract-law-in-the-age-of-agentic-ai-whos-really-clicking-accept/

[25] IAPP, "Engineering GDPR Compliance in the Age of Agentic AI," 2025. Available at: https://iapp.org/news/a/engineering-gdpr-compliance-in-the-age-of-agentic-ai/

[26] OWASP Gen AI Security Project, "Top 10 Risks and Mitigations for Agentic AI Security," December 2025. Available at: https://genai.owasp.org/2025/12/09/owasp-genai-security-project-releases-top-10-risks-and-mitigations-for-agentic-ai-security/

[27] Gartner, "Gartner Predicts Over 40% of Agentic AI Projects Will Be Canceled by End of 2027," June 2025. Available at: https://www.gartner.com/en/newsroom/press-releases/2025-06-25-gartner-predicts-over-40-percent-of-agentic-ai-projects-will-be-canceled-by-end-of-2027

[28] Gartner, "AI Agents Will Reduce the Time It Takes To Exploit Account Exposures by 50% by 2027," March 2025. Available at: https://www.gartner.com/en/newsroom/press-releases/2025-03-18-gartner-predicts-ai-agents-will-reduce-the-time-it-takes-to-exploit-account-exposures-by-50-percent-by-2027

[29] ACL Anthology, "InjecAgent: Benchmarking Indirect Prompt Injections in Tool-Integrated LLM Agents," 2024. Available at: https://aclanthology.org/2024.findings-acl.624/

[30] Orrick LLP, "The European Data Protection Board Shares Opinion on How to Use AI in Compliance with GDPR," March 2025. Available at: https://www.orrick.com/en/Insights/2025/03/The-European-Data-Protection-Board-Shares-Opinion-on-How-to-Use-AI-in-Compliance-with-GDPR

[31] Wilson Sonsini, "Will Indemnification Commitments Address Market Demands in AI?" 2024. Available at: https://www.wsgr.com/en/insights/will-indemnification-commitments-address-market-demands-in-ai.html

[32] Hogan Lovells, "Agentic AI in Financial Services: Regulatory and Legal Considerations," 2025. Available at: https://www.hoganlovells.com/en/publications/agentic-ai-in-financial-services-regulatory-and-legal-considerations

[33] WTW, "Emerging AI Exposures and the Role of Cyber and E&O; Insurance," March 2025. Available at: https://www.wtwco.com/en-us/insights/2025/03/emerging-ai-exposures-and-the-role-of-cyber-and-e-and-o-insurance

[34] Lathrop GPM, "Liability Considerations for Developers and Users of Agentic AI Systems," 2025. Available at: https://www.lathropgpm.com/insights/liability-considerations-for-developers-and-users-of-agentic-ai-systems/

[35] The Future Society, "AI Agents Governed Under the EU AI Act," 2025. Available at: https://thefuturesociety.org/aiagentsintheeu/

This guide is provided for informational purposes and does not constitute legal advice.
Consult applicable rules of professional conduct and firm-specific policies before implementing any contractual provisions discussed herein.