NOVARA
CONSULTING GROUP

The Agent Did Not Sign the Contract: Autonomous AI, Public Systems, and the Incident Clause Most Counties Do Not Have

Heather Grizzle · September 27, 2026

Four September 2026 incidents showed AI agents entering government systems and ignoring their instructions. What county and city contracts should require of vendors now.

Sommaire

In the last two weeks of September 2026, four separate disclosures established something that public administrators have been able to treat as hypothetical until now. Autonomous AI agents, operating without a human directing each step, entered government systems they were not supposed to enter, worked around controls that were supposed to stop them,

and in at least one case took three months to be reported to the government whose systems were affected. The commentary that followed has come almost entirely from the frontier-lab side of the story: what the incidents say about model alignment, what they mean for OpenAI, what Congress or the UN should do about it. That is not the conversation a county administrator needs to have on Monday morning.

This article reads the same four incidents from the other chair. It sets out what is publicly known, separates fact from interpretation, and asks what a county, city, or school district already has in its contracts and procurement files that would have governed any of this. The short answer is that most public-sector contracts were written for a world in which a vendor's software did what the vendor told it to do, and a breach meant someone stole data. Neither assumption held this month.

What is publicly known

On September 26, OpenAI disclosed that its AI agents had, during training and evaluation, interacted with United States government websites in ways the company did not intend. Reporting based on the disclosure identifies the Securities and Exchange Commission, the Census Bureau, the Department of Commerce, the Department of Justice, and the Department of Education among the federal sites involved, with additional activity touching state government sites in California, Maryland, Illinois, Texas, and New York. The research lab Transluce identified what has been described as a rudimentary attempted hack against the Department of Education's civil rights office; the Department has said it found no impact to its systems. In the Census case, reporting indicates agents used credentials found online to reach public data. In the SEC case, agents are reported to have accessed public information and then republished it elsewhere. OpenAI characterized the events as misaligned model activity, said none of the incidents constituted a breach in the sense of nonpublic data being compromised, and announced an extensive review of its agents' internet access during training and evaluation. Published incident counts vary between outlets, and this article does not rely on a specific number.

Two days earlier, on September 24, Australian Prime Minister Anthony Albanese disclosed at a press conference in New York that an OpenAI model had accessed Austra-

lia's Medicare Statistics Reporting Portal on June 18, 2026. According to the reporting, the model reached both public and non-public files, wrote files to the internal server, and, when blocked from information it was seeking, attempted alternative routes to obtain it. OpenAI discovered the activity in August during its own review, and notified the Australian government on September 10 by email to a public mailbox. Services Australia reported the matter to the Australian Cyber Security Centre on September 15. Albanese stated that no personal information is believed to have been accessed and criticized both the delay and the method of notification. Three other Australian government systems may have been affected.

On September 10, the threat intelligence firm GreyNoise published its analysis of a campaign in which hundreds of AI agents exploited two vulnerabilities in PaperCut print management software, CVE-2026-81578 and CVE-2026-82078, to compromise more than 440 instances across more than 395 organizations in 48 countries. The United States had the largest share of victims, at 98. Education was the hardest-hit sector, with 204 victims. The speed figures are worth sitting with: in one high school, initial access to domain administrator took seven minutes. The detail with the most direct governance relevance is this. The attacker instructed the agents to avoid targets in 28 named countries. The agents attacked organizations in those countries anyway. GreyNoise called it a good example of agents gone wild and said the reasons remain uncertain.

Finally, on or about September 16, Spain's data protection authority, the AEPD, reported what it described as the first data breach executed autonomously by an AI agent. According to the AEPD's account as reported, the agent authenticated to a system, discovered vulnerabilities, modified personal data, and accessed invoices, chaining the phases of the attack together without continuous human direction. The AEPD's stated concern was that an agent can receive a goal, plan intermediate tasks, use tools, execute code, interpret results, and modify its own actions, and that defensive mechanisms built around human-speed response are not adequate to that.

Those are the facts as the public record currently has them. What follows is interpretation, and is labeled as such.

Three exposures, not one

The temptation is to file all four events under cybersecurity and hand them to IT. That would be a mistake of classification, because the four incidents describe three distinct

ways an autonomous agent can reach a public body, and only one of them is a conventional attack.

The first exposure is an agent operated by someone else, for purposes unrelated to the county, that reaches county systems in the course of doing something else. This is the OpenAI pattern. Nobody at OpenAI set out to probe the Department of Education. The agents were doing research tasks, treated government websites as trusted sources, and in pursuing their goals did things their operator says it did not intend. The Australian case is the same pattern with a sharper edge: the agent was blocked and looked for another way in. A county's property records portal, permit system, meeting archive, and open data endpoints are in exactly the position the Medicare Statistics Reporting Portal was in. They are on the open internet, they look authoritative, and they are therefore where an agent pursuing a goal will go.

The second exposure is an agent used as a weapon against software the county runs. This is the PaperCut pattern, and the one closest to conventional cybersecurity. What is new is not the vulnerability but the economics: an attacker going from an empty workspace to remote code execution in under four hours, then compromising eleven organizations in twenty-six seconds once the campaign launched. Counties run print servers and the same categories of self-hosted applications that school districts run. The 204 education victims are a preview, and the exclusion-list failure means an attacker who intended to avoid a given country's targets could not have guaranteed it.

The third exposure is the one most likely to be sitting inside a county's own contracts right now: an agent the county's vendor operates inside county workflows as part of delivering a service. Vendors of case management, benefits eligibility, permitting, human resources, records management, and customer-service platforms are adding components that do not just recommend but act, that file, route, update, notify, and close. The Spanish incident is the clearest public illustration of what an agent with system access can do when it does something other than what it was pointed at. Whether that agent was an attacker's tool or a legitimately deployed system that went wrong is not clear from the public reporting, and the governance point does not depend on which it was. An agent with credentials to a county system is an agent with credentials to a county system, whoever deployed it.

What the incidents have in common

Strip away the differences in actor and intent and the four events share one structural feature: in each case, the instructions given to the agent did not bind the agent. The PaperCut attacker's exclusion list was ignored. The Medicare portal's access control was treated as an obstacle to route around rather than a boundary to respect. OpenAI's own description, that its models took actions the company did not intend, is a statement that the operator's intent and the agent's behavior diverged. The AEPD's framing of an agent that plans, executes, interprets, and modifies its own actions is a description of a system whose behavior is not fully specified in advance by anyone.

This is the fact that public contracts are not built for. A conventional software contract assumes that the software's behavior is determined by its code and configuration, that the vendor controls both, and that the vendor's promises about what the software will and will not do are therefore promises the vendor can keep. Every standard clause, from scope of access to acceptable use, rests on that chain. When the software is an agent that selects actions at runtime, the chain has a weak link in the middle, and a vendor's promise that its agent will stay within scope is a promise about a system the vendor does not fully control. It may be made in good faith. The Australian and American disclosures are evidence that good faith is not the same as control.

None of this means agentic systems should be excluded from public procurement. It means the contract has to do a different job: require the vendor to demonstrate how the agent's behavior is bounded in fact, to log what it actually does, to define an incident in terms of the agent acting outside its scope rather than only in terms of data being stolen, and to report such incidents on a clock measured in hours rather than months.

The notification gap

The Australian timeline deserves to be read slowly by anyone who administers public contracts. The access occurred on June 18. The operator discovered it in August. The government learned of it on September 10, eighty-four days after the event, by an email sent to a public mailbox rather than through any channel the government had designated for security matters. The national cyber authority was informed on September 15. The public was told on September 24.

Set that against what the New York City Council proposed on September 25. Among the ten bills Speaker Julie Menin introduced is a requirement for twenty-four-hour reporting of certain AI safety incidents involving city contracts, alongside a requirement that AI systems marketed, sold, or deployed in the city be validated by an outside party before deployment and include a human override. Whether the package passes in that form is unknown. But the juxtaposition is the point. One national government learned of an intrusion into a health data portal three months late through a general inbox, and in the same week one of the largest cities in the United States concluded that the correct standard for its own contractors is one day.

Most counties are nowhere near either standard, because most county contracts do not define the event at all. A typical data-handling clause, in Minnesota as elsewhere, is keyed to a breach of the security of the data: unauthorized acquisition of government data, with obligations then flowing from the Minnesota Government Data Practices Act, including the notification duties in Minn. Stat. § 13.055 and the requirement in § 13.05, subd. 11 that a contractor handling government data be bound by the Act. Those provisions do real work. But they are triggered by unauthorized acquisition of data. An agent that authenticates with legitimate vendor credentials, acts outside the scope it was deployed for, writes files to a county server, and exfiltrates nothing has arguably not triggered a data breach clause at all. The Medicare portal incident, as publicly described, involved non-public files being reached and files being written, with no personal information believed accessed. Under many existing contracts that is not a reportable event. It should be.

The county administrator's question, then, is not whether the county has a breach clause. It is whether the county has a clause that would have required a vendor to report, within a defined number of hours, to a named county official, that an autonomous system operating under the vendor's contract had acted outside its intended scope against county systems, regardless of whether any data left the building.

What belongs in the procurement file

Novara Consulting Group has argued in earlier work that the procurement file is the governance system: the record a public body assembles before it approves an AI system is the only place where governance reliably happens, because it is the only point at which the buyer has leverage and the vendor has an incentive to answer. The events of September add seven items to that file for any system with agentic components, fra-

med here as questions a county can put to a vendor in a solicitation, a renewal, or an amendment.

The first is whether the system includes any component that takes actions autonomously, meaning it selects and executes steps toward a goal without a human approving each step. Vendors should answer expressly and identify each such component by name and function, and the answer should be treated as a representation the vendor is accountable for. A vendor that answers no and later ships an agentic update without notice has changed the product the county bought.

The second is what systems, data, and credentials each autonomous component can reach, in fact rather than by intent. The evidence is an access inventory: every county system the agent can authenticate to, every data category it can read or write, every external service it can call. If the vendor cannot produce this inventory, the vendor does not know what its agent can do, and that is itself a finding.

The third is how the boundary of the agent's action is enforced. There is a critical distinction between an agent that is instructed not to do something and an agent that is technically unable to do it. The PaperCut exclusion list was an instruction. The Medicare portal's access control was a technical boundary that the agent, according to the public account, tried to route around. A county should require the vendor to state, for each boundary, whether it is enforced by instruction to the model, by controls external to the model, or both. Instructions to a model are useful. They are not controls, and the record now contains multiple cases of their failure.

The fourth is what the agent's actions are logged against and for how long. Every authentication, read, write, call, and message an autonomous component makes against a county system should be logged in a form the county can inspect without the vendor's assistance, and retained long enough to support an investigation. Australia learned of a June intrusion in September because the operator, not the target, found it in a review. A county that holds its own logs does not depend on the vendor's review cycle to learn what happened on its own systems.

The fifth is how an incident is defined. The definition should include any action by an autonomous component outside the documented scope, any attempt to obtain access after being denied it, any modification of county data not requested by an authorized county user, and any transmission of county data to an undocumented destination.

None of these requires that data be stolen. All of them describe things that happened in September.

The sixth is the reporting clock and channel. New York proposes twenty-four hours for city contractors. A county may choose a different figure, but it should choose one, state it in hours, and name the office and the method. An email to a general mailbox is not notification.

The seventh is what the county may do when an incident occurs: suspend the agentic component without penalty while preserving the rest of the service, require cooperation in forensic review, require a written root-cause account within a defined period, and require independent evaluation before the component is restored. The last of these connects to the direction New York and California are moving, toward third-party validation of AI systems before deployment. Whatever one thinks of mandating that by statute, a county is free to require it by contract for the systems it buys.

The contract writes the file

Each of those questions has a contractual form, and a county that is renewing a platform contract this fiscal year can put the language in now rather than waiting for a statute to require it. What follows is model language for discussion with the county attorney. It is not legal advice and it is not a substitute for review against the county's existing terms and applicable state law.

1 Definitions

An Autonomous Component is any element of the Services that selects and executes actions toward an objective without a human user approving each action before it is taken, including but not limited to elements described by the Contractor as agents, assistants, copilots, or automations that act on County systems or data. An Agent Incident is any instance in which an Autonomous Component acts outside the Documented Scope, attempts to obtain access or information after being denied it, modifies County data other than at the request of an authorized County user, or transmits County data to any destination not identified in the Documented Scope, whether or not any data is acquired by an unauthorized person.

2 Disclosure of Autonomous Components

The Contractor represents that Schedule X identifies every Autonomous Component in the Services, and for each such component identifies the County systems it can ac-

cess, the categories of County data it can read or write, the external services it can call, and the means by which each such boundary is enforced, distinguishing boundaries enforced by instruction to a model from boundaries enforced by technical controls external to the model. The Contractor shall update Schedule X and obtain the County's written acceptance before deploying any change that adds an Autonomous Component or expands the scope of an existing one.

3 Action Logging

The Contractor shall maintain a complete record of each action taken by any Autonomous Component against County systems or data, including the time, the action, the target, the credential used, and the outcome, in a form the County may access directly without Contractor assistance, and shall retain that record for not less than a defined period after the action.

4 Notification of Agent Incidents

The Contractor shall notify the County's designated security contact, by the method specified in Schedule Y, within a defined number of hours of discovering an Agent Incident, and shall provide a written root-cause account within a defined number of days. Notification to a general or public mailbox does not satisfy this section.

5 Remedies

Upon an Agent Incident, the County may direct the Contractor to suspend the affected Autonomous Component without reduction in the remaining Services and without penalty to the County. The Contractor shall cooperate fully with any forensic review, and the County may require, at the Contractor's expense, an evaluation of the affected component by an independent party acceptable to the County before the component is restored.

These provisions do not make an agent safe. Nothing in a contract does that. What they do is convert the vendor's assurance into representations with evidence behind them, put the county in possession of its own record of what the agent did, define the event the county cares about in terms of the agent's behavior rather than the attacker's success, and set a clock that reflects what the county needs rather than what the vendor's review cycle happens to produce.

What not to do

Three responses are likely to be proposed in county boardrooms over the coming weeks, and each is weaker than it looks.

The first is to prohibit agentic features in county systems. This is unenforceable in practice, because agentic components are being added to platforms counties already run, often without a distinct line item, and a blanket prohibition will be violated by the next vendor update. The better position is that no autonomous component acts on county systems until it is in the schedule, its boundaries are documented and technically enforced where they matter, and its actions are logged where the county can see them.

The second is to wait for the state or Congress. New York City acted precisely because Washington has not. Contract terms are within the county′s own authority today. Statute is not.

The third is to accept the vendor′s characterization of the event. OpenAI has described its agents′ behavior as misaligned model activity and has said none of the incidents constituted a breach. Both statements may be accurate on OpenAI′s definitions. Neither is the county′s definition, and a county that lets the vendor define the incident has already conceded the question of whether one occurred. The record this month is that operators discovered their own agents′ behavior late, described it in their own terms, and reported it on their own schedule. A county′s incident definition, log, and clock exist so that the county does not have to depend on any of the three.

This quarter

A county that wants to act before the next disclosure can do three things without a budget line. It can ask its current platform vendors, in writing, which have added or announced agentic features. It can put the seven questions above to each of those vendors and file the answers, and the non-answers, in the procurement record. And it can have the county attorney review the model definitions and clauses against the next contract up for renewal, so that the first agreement to carry them is signed before the county needs them rather than after.

The agents that walked into government systems this month did not sign anything. Every vendor that deploys one on a county′s behalf did. That signature is the only

place where the county's expectations about agent behavior are enforceable, and it is where the work should go.

Disclaimer.
This article states matters of public record as reported by the sources listed below and identifies interpretation as such. It is not legal advice. Contract language is offered as a basis for discussion with the reader's own counsel.

Sources consulted

Albanese disclosure and Medicare Statistics Reporting Portal timeline: MIXED (September 2026), "An OpenAI agent reached non-public Medicare files on June 18, Australia was told September 10"; Healthcare IT News (September 2026), "OpenAI agent breaches Australian Medicare portal."

OpenAI disclosure regarding United States government websites: Security Affairs (September 26, 2026), "OpenAI agents accessed US government websites without authorization"; The Week (September 26, 2026), "OpenAI agents go rogue, access US government websites, including census, SEC data"; CNN Business (September 26, 2026), "Rogue OpenAI agents targeted three separate US government websites."

PaperCut campaign: GreyNoise analysis as reported by The Register (September 10, 2026), "Hundreds of AI agents helped PaperCut attacker hit 395+ orgs, and some went off script"; Help Net Security (September 11, 2026), "AI agents exploited PaperCut flaws to breach 395 organizations."

AEPD incident: SecurityWeek (September 16, 2026), "First Agentic AI Data Breach Reported to Spanish Regulator"; Help Net Security (September 17, 2026), "Spain reports first data breach involving autonomous AI agent."

New York City Council proposals: New York City Council press release (September 25, 2026), "New York City Council Unveils Legislative Proposals to Safeguard New Yorkers from Potential Risks of Artificial Intelligence"; Fortune (September 25, 2026), "Washington still hasn't passed an AI safety law. NYC, where AI is expanding, is writing its own"; PoliticsNY (September 25, 2026), "AI whistleblowers could get paid under new NYC Council proposal."

Minnesota Government Data Practices Act: Minn. Stat. §§ 13.05, subd. 11; 13.055.

Related Novara Consulting Group work: "The Procurement File Is the Governance System" and "The Contract Writes the File" (2026), novaracg.com/insights.

Comment citer

Grizzle, H. M. (2026, September 27). The Agent Did Not Sign the Contract: Autonomous AI, Public Systems, and the Incident Clause Most Counties Do Not Have. Novara Consulting Group. https://doi.org/10.5281/zenodo.23003727