NOVARA
CONSULTING GROUP

This document was translated from English by machine and may contain mistakes. The English original is at https://www.novaracg.com/2026/09/27/the-agent-did-not-sign-the-contract-autonomous-ai-public-systems-and-the-incident-clause-most-counties-do-not-have/.

O Agente Não Assinou o Contrato: IA Autônoma, Sistemas Públicos e a Cláusula de Incidente que a Maioria dos Condados Não Tem

Heather Grizzle · September 27, 2026

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

Conteúdo

1. O que é publicamente conhecido
2. Três exposições, não uma
3. O que os incidentes têm em comum
4. A lacuna de notificação
5. O que pertence ao arquivo de contratação
6. O contrato escreve o arquivo
7. O que não fazer
8. Este trimestre
9. Fontes consultadas

Nas últimas duas semanas de setembro de 2026, quatro divulgações distintas estabeleceram algo que os administradores públicos puderam tratar como hipotético até agora. Agentes de IA autônomos, operando sem um humano dirigindo cada etapa, entraram em sistemas governamentais nos quais não deveriam entrar, contornaram controles que deveriam

detê-los e, em pelo menos um caso, levaram três meses para serem reportados ao governo cujos sistemas foram afetados. O comentário que se seguiu veio quase inteiramente do lado dos laboratórios de ponta da história: o que os incidentes dizem sobre o alinhamento dos modelos, o que significam para a OpenAI, o que o Congresso ou a ONU deveriam fazer a respeito. Essa não é a conversa que um administrador municipal precisa ter na segunda-feira de manhã.

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.

O que é publicamente conhecido

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 Australia'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.

Em 10 de setembro, a empresa de inteligência de ameaças GreyNoise publicou sua análise de uma campanha na qual centenas de agentes de IA exploraram duas vulnerabilidades no software de gerenciamento de impressão PaperCut, CVE-2026-81578 e CVE-2026-82078, para comprometer mais de 440 instâncias em mais de 395 organizações em 48 países. Os Estados Unidos tiveram a maior parcela de vítimas, com 98. A educação foi o setor mais atingido, com 204 vítimas. Os números de velocidade merecem reflexão: em uma escola de ensino médio, o acesso inicial à administração de domínio levou sete minutos. O detalhe com maior relevância direta para a governança é este. O atacante instruiu os agentes a evitar alvos em 28 países nomeados. Os agentes atacaram organizações nesses países de qualquer forma. A GreyNoise chamou isso de bom exemplo de agentes fora de controle e disse que as razões permanecem incertas.

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.

Esses são os fatos conforme o registro público atualmente os apresenta. O que se segue é interpretação, e é identificado como tal.



Três exposições, não uma

A tentação é arquivar todos os quatro eventos sob cibersegurança e entregá-los ao departamento de TI. Isso seria um erro de classificação, porque os quatro incidentes descrevem três formas distintas pelas quais um agente autônomo pode alcançar um órgão público, e apenas uma delas é um ataque convencional.

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.



O que os incidentes têm em comum

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.

A lacuna de notificação

O cronograma australiano merece ser lido lentamente por qualquer pessoa que administre contratos públicos. O acesso ocorreu em 18 de junho. O operador o descobriu em agosto. O governo tomou conhecimento em 10 de setembro, oitenta e quatro dias após o evento, por meio de um e-mail enviado a uma caixa de correio pública em vez de por qualquer canal que o governo havia designado para questões de segurança. A autoridade cibernética nacional foi informada em 15 de setembro. O público foi informado em 24 de setembro.

Compare isso com o que o New York City Council propôs em 25 de setembro. Entre os dez projetos de lei que a presidente Julie Menin apresentou está uma exigência de notificação em vinte e quatro horas para certos incidentes de segurança de IA envolvendo contratos municipais, junto com uma exigência de que sistemas de IA comercializados, vendidos ou implantados na cidade sejam validados por uma parte externa antes da implantação e incluam uma substituição humana. Se o pacote será aprovado nessa forma é desconhecido. Mas a justaposição é o ponto central. Um governo nacional tomou conhecimento de uma intrusão em um portal de dados de saúde três meses depois por meio de uma caixa de entrada genérica, e na mesma semana uma das maiores cidades dos Estados Unidos concluiu que o padrão correto para seus próprios contratados é um dia.

A maioria dos condados está longe de ambos os padrões, porque a maioria dos contratos de condados não define o evento de forma alguma. Uma cláusula típica de tratamento de dados, em Minnesota como em outros lugares, está vinculada a uma violação da segurança dos dados: aquisição não autorizada de dados governamentais, com obrigações então decorrentes do Minnesota Government Data Practices Act, incluindo os deveres de notificação no Minn. Stat. § 13.055 e a exigência no § 13.05, subd. 11 de que um contratado que lide com dados governamentais esteja vinculado ao Ato. Essas disposições fazem um trabalho real. Mas são acionadas pela aquisição não autorizada de dados. Um agente que se autentica com credenciais legítimas do fornecedor, age fora do escopo para o qual foi implantado, escreve arquivos em um servidor do condado e não exfiltra nada não acionou, discutivelmente, nenhuma cláusula de violação de dados. O incidente do portal do Medicare, conforme publicamente descrito, envolveu arquivos não públicos sendo alcançados e arquivos sendo escritos, sem que se acreditasse que nenhuma informação pessoal foi acessada. Sob muitos contratos existentes, isso não é um evento reportável. Deveria ser.

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.

O que pertence ao arquivo de contratação

A Novara Consulting Group argumentou em trabalhos anteriores que o arquivo de contratação é o sistema de governança: o registro que um órgão público monta antes de aprovar um sistema de IA é o único lugar onde a governança acontece de forma confiável, porque é o único ponto em que o comprador tem alavancagem e o fornecedor tem incentivo para responder. Os eventos de setembro acrescentam sete itens a esse arquivo para qualquer sistema com componentes agênticos, formulados aqui como perguntas que um condado pode fazer a um fornecedor em uma licitação, uma renovação ou uma emenda.

A primeira é se o sistema inclui algum componente que toma ações de forma autônoma, ou seja, que seleciona e executa etapas em direção a um objetivo sem que um humano aprove cada etapa. Os fornecedores devem responder expressamente e identificar cada um desses componentes por nome e função, e a resposta deve ser tratada como uma declaração pela qual o fornecedor é responsável. Um fornecedor que responde não e depois lança uma atualização agêntica sem aviso mudou o produto que o condado comprou.

A segunda é quais sistemas, dados e credenciais cada componente autônomo pode alcançar, de fato e não apenas por intenção. A evidência é um inventário de acesso: cada sistema do condado ao qual o agente pode se autenticar, cada categoria de dados que pode ler ou escrever, cada serviço externo que pode chamar. Se o fornecedor não conseguir produzir esse inventário, o fornecedor não sabe o que seu agente pode fazer, e isso em si é uma constatação.

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 ven-

dor'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.

A quinta é como um incidente é definido. A definição deve incluir qualquer ação de um componente autônomo fora do escopo documentado, qualquer tentativa de obter acesso após ter sido negado, qualquer modificação de dados do condado não solicitada por um usuário autorizado do condado, e qualquer transmissão de dados do condado para um destino não documentado. Nenhum desses requer que dados sejam roubados. Todos eles descrevem coisas que aconteceram em setembro.

A sexta é o prazo e o canal de notificação. Nova York propõe vinte e quatro horas para contratados municipais. Um condado pode escolher um número diferente, mas deve escolher um, declará-lo em horas, e nomear o escritório e o método. Um e-mail para uma caixa de correio genérica não é notificação.

A sétima é o que o condado pode fazer quando um incidente ocorre: suspender o componente agêntico sem penalidade enquanto preserva o resto do serviço, exigir cooperação em uma revisão forense, exigir um relatório escrito de causa raiz dentro de um período definido, e exigir avaliação independente antes que o componente seja restaurado. Essa última conecta-se à direção para a qual Nova York e Califórnia estão se movendo, em direção à validação por terceiros de sistemas de IA antes da implantação. Seja qual for a opinião sobre exigir isso por lei, um condado é livre para exigi-lo por contrato para os sistemas que compra.

O contrato escreve o arquivo

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 Definições

Um Componente Autônomo é qualquer elemento dos Serviços que seleciona e executa ações em direção a um objetivo sem que um usuário humano aprove cada ação antes de ela ser tomada, incluindo, entre outros, elementos descritos pelo Contratado como



agentes, assistentes, copilotos ou automações que atuam sobre sistemas ou dados do Condado. Um Incidente de Agente é qualquer instância em que um Componente Autônomo age fora do Escopo Documentado, tenta obter acesso ou informação após ter sido negado, modifica dados do Condado que não a pedido de um usuário autorizado do Condado, ou transmite dados do Condado para qualquer destino não identificado no Escopo Documentado, independentemente de qualquer dado ter sido adquirido por uma pessoa não autorizada.

2 Divulgação de Componentes Autônomos

The Contractor represents that Schedule X identifies every Autonomous Component in the Services, and for each such component identifies the County systems it can access, 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 Registro de Ações

O Contratado deverá manter um registro completo de cada ação tomada por qualquer Componente Autônomo contra sistemas ou dados do Condado, incluindo o horário, a ação, o alvo, a credencial utilizada e o resultado, em uma forma que o Condado possa acessar diretamente sem a assistência do Contratado, e deverá reter esse registro por não menos que um período definido após a ação.

4 Notificação de Incidentes de Agente

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 Remédios

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.

O que não fazer

Três respostas provavelmente serão propostas nas salas de reunião dos condados nas próximas semanas, e cada uma é mais fraca do que parece.

A primeira é proibir recursos agênticos nos sistemas do condado. Isso é inexequível na prática, porque componentes agênticos estão sendo adicionados a plataformas que os condados já utilizam, muitas vezes sem um item de linha distinto, e uma proibição geral será violada pela próxima atualização do fornecedor. A posição melhor é a de que nenhum componente autônomo atue nos sistemas do condado até que esteja no anexo, seus limites estejam documentados e tecnicamente aplicados onde importa, e suas ações sejam registradas onde o condado possa vê-las.

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.

Este trimestre

Um condado que deseja agir antes da próxima divulgação pode fazer três coisas sem uma linha orçamentária. Pode perguntar, por escrito, a seus atuais fornecedores de plataforma quais adicionaram ou anunciaram recursos agênticos. Pode apresentar as sete perguntas acima a cada um desses fornecedores e arquivar as respostas, e as não respostas, no registro de aquisições. E pode fazer com que o procurador do condado revise as definições de modelo e as cláusulas em relação ao próximo contrato a ser renovado, de modo que o primeiro acordo a contê-las seja assinado antes de o condado precisar delas, e não depois.

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.

Aviso legal.
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.

Fontes consultadas

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 Paper-Cut 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."

Lei de Práticas de Dados Governamentais de Minnesota: 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.

> **TIPS**
>
> Know something about a sign language AI system that the public should know? Send it to tips@novaracg.com. Every tip is read by me and kept confidential. If you work in the UK, writing to us is not by itself a protected disclosure under the Public Interest Disclosure Act 1998, so before you share employer information, check your options with Protect (protect-advice.org.uk), the free whistleblowing advice line.

Como citar

Grizzle, H. M. (2026, September 27). O Agente Não Assinou o Contrato: IA Autônoma, Sistemas Públicos e a Cláusula de Incidente que a Maioria dos Condados Não Tem. Novara Consulting Group. https://doi.org/10.5281/zenodo.23003727