Notion AI : qui garde la promesse de non-conservation ?

Début novembre 2026, des transcriptions de conversations avec Notion AI ont été atteintes chez Fireworks AI, le prestataire qui les analysait. L'affaire a été racontée comme une attaque en chaîne contre un sous-traitant de l'intelligence artificielle. Elle montre autre chose. La clause de « zero data retention » (conservation nulle des données) est un gardien de la confiance impersonnelle : un dispositif qui remplace l'observation que le client ne peut pas faire. Comme tous les gardiens, elle exige à son tour qu'on lui fasse confiance.

Par Alexandre Berge · 19 novembre 2026 · 13 min de lecture · Chronique

Une attaque en chaîne, telle qu'elle a été racontée

Le 2 novembre au soir, Fireworks AI prévient Notion qu'un petit jeu de données a été consulté par des tiers malveillants. Notion l'identifie comme des données de test synthétiques de son assistant, sans donnée client. Le lendemain matin, par précaution, l'éditeur coupe ses connexions avec le prestataire et fait supprimer ses données. Selon BleepingComputer, il arrête aussi les entraînements qu'il y menait. Sa page d'aide date du 4 novembre la suppression complète. Le 6 novembre, Fireworks signale que des données de clients ont pu être atteintes. Le 8, il confirme l'existence de ce second jeu de données.

Le jeu contient des transcriptions de conversations avec Notion AI, utilisées par la fonction Insights. Selon Notion, moins de 1 % des espaces de travail sont concernés. Dans l'immense majorité des cas, les conversations ne portent ni le nom de leur auteur ni celui de l'espace. Elles contiennent pourtant le texte intégral des messages et des réponses, y compris les mots de passe ou les clés d'API qu'un utilisateur aurait collés. L'éditeur demande à ses clients de traiter ces données comme exposées. Il a suspendu Insights et cessé de faire servir des modèles par Fireworks.

Côté Fireworks, l'affaire avait commencé plus tôt. Le bulletin du 9 octobre décrit l'usage non autorisé d'identifiants internes. Cet accès a permis de lire les variables d'environnement de tâches d'évaluation, et donc les clés et jetons qui y étaient stockés, dans un nombre limité de comptes clients. Le même bulletin précise que Fireworks ne conserve par défaut ni les requêtes ni les réponses, au motif que « our platform is ZDR » (« notre plateforme est à conservation nulle »). La mise à jour du 9 novembre élargit nettement le périmètre. Le 23

octobre, un fournisseur cloud tiers qui hébergeait les machines de recherche de Fireworks a été compromis. Le 24, l'attaquant a atteint ses machines de développement. Sont touchés des métadonnées, des informations de facturation et, pour certains clients, des jeux de données et des poids de modèles incomplets (les paramètres appris qui constituent un modèle). Fireworks attribue l'opération à un groupe criminel qui revend des accès détournés à des modèles de langage (le « LLMjacking »).

The Register, le 10 novembre, et BleepingComputer, le lendemain, racontent l'affaire sous cet angle : un prestataire d'inférence (le service qui fait tourner les modèles) atteint deux fois, un hébergeur compromis, des données de clients qui fuient par ricochet. Les faits établis s'arrêtent au 15 novembre 2026. Ce récit est exact. Il laisse pourtant de côté une phrase de Notion.

La conservation découverte dans le butin, pas par un contrôle

Dans la page d'aide consacrée à l'incident, Notion présente la conservation nulle comme un engagement envers ses clients Enterprise. Elle porte sur les requêtes envoyées aux fournisseurs de modèles, dont Notion attend qu'ils ne gardent rien. La page ajoute que Fireworks s'était engagé contractuellement à la conservation nulle pour l'usage que Notion faisait de sa plateforme. Puis vient le constat : « data was retained from the Fireworks endpoints that we used » (« des données ont été conservées depuis les points d'accès de Fireworks que nous utilisions »).

La chronologie dit comment cette conservation est devenue visible. Le 3 novembre, Notion se déconnecte en croyant que seules des données synthétiques sont en cause. Ce sont les découvertes ultérieures de Fireworks qui révèlent l'existence de transcriptions réelles. Aucun contrôle du client n'a constaté l'écart entre la clause et la pratique. **Cet écart est apparu quand le butin d'un attaquant a été inventorié.**

Le récit de l'attaque en chaîne désigne un maillon faible : un hébergeur, des identifiants, un groupe criminel. Il passe sous silence ce qui fondait la relation entre Notion et Fireworks. Notion n'observait pas ce que faisait son prestataire. Il s'en remettait à une clause.

Une confiance qui ne repose sur aucune observation

> **La confiance impersonnelle (Susan P. Shapiro, 1987)**
>
> Dans « The Social Control of Impersonal Trust » (*American Journal of Sociology*, 1987), la sociologue Susan Shapiro décrit la confiance comme une relation d'agence : un mandant confie des ressources ou une responsabilité à un agent qui agit pour lui. La confiance devient impersonnelle lorsque aucun lien direct ne permet au mandant de surveiller, d'évaluer ou de contraindre l'agent. Les sociétés y répondent par des « gardiens de la confiance » : normes de procédure, obligations d'information, séparation des fonctions,

> licences, contrôles, audits, assurances. Le paradoxe central est que ces gardiens sont eux-mêmes des agents. Ils ont leur propre expertise, leur propre marge de manœuvre et leurs propres occasions de défaillance. Chaque couche de protection appelle donc une confiance nouvelle, que d'autres gardiens doivent à leur tour encadrer.

La relation entre Notion et Fireworks correspond trait pour trait à la confiance impersonnelle. Notion ne voit pas les machines de son prestataire. Il ne peut pas constater qu'une requête traitée a été effacée. La conservation nulle est d'ailleurs une qualité que l'usage ne révèle jamais. Le service répond de la même manière, que les données soient gardées ou non. Une chronique précédente l'a montré à propos de la cause des pannes de logiciel en ligne (« SaaS : la cause d'une panne, un bien de croyance », 17 octobre 2026). Ce qui est nouveau ici, c'est la forme que prend la protection du client.

Shapiro aide à lire cette forme. La clause contractuelle est une *norme de procédure*. Le bulletin du 9 octobre, qui affirmait la conservation nulle, est une *obligation d'information*. Shapiro observe que les règles d'information encadrent ce que l'agent déclare sans garantir que la déclaration soit vraie. Les deux gardiens étaient en place. Aucun ne donnait au client un moyen d'observer. Et la chaîne ne s'arrête pas là. Le client Enterprise fait confiance à Notion, qui fait confiance à Fireworks, qui faisait confiance à l'hébergeur de ses machines de recherche. **Chaque maillon garantit au suivant ce qu'il ne peut lui-même que croire.**

Le stock exposé venait d'un outil de contrôle de l'usage

Le second fait porte sur l'origine des données exposées. Notion décrit Insights comme un outil qui donne aux administrateurs des informations sur la façon dont les membres de leur espace utilisent l'IA. L'onglet est réservé aux propriétaires et aux administrateurs des espaces des offres Business et Enterprise. Pour produire ces analyses, Notion explique avoir envoyé à Fireworks les requêtes et les réponses stockées de ses utilisateurs.

La même page précise le périmètre de la promesse faite aux clients. La conservation nulle couvre les requêtes envoyées aux fournisseurs de modèles. Elle ne vise ni l'historique des conversations ni Insights, que Notion doit stocker pour rendre ces services. Deux engagements se superposent donc. Le premier, de Fireworks à Notion, couvre l'usage de la plateforme. Le second, de Notion à ses clients, exclut précisément la fonction dont les transcriptions ont été exposées. Le client final ne voit que le second. Il n'a aucun moyen de savoir quelles fonctions de son espace envoient des copies de conversations à quel prestataire.

Shapiro avait relevé que la recherche de fidélité des agents crée ses propres problèmes. Le cas Notion en donne une version d'entreprise. **L'outil conçu pour que la direction voie comment ses salariés utilisent l'IA a constitué la copie qui s'est retrouvée hors de son contrôle.** Ce constat porte sur l'architecture des engagements et ne prête aucune intention à

quiconque.

À la date du 15 novembre 2026, ni les deux pages de Notion ni les deux bulletins de Fireworks ne disent par quelle voie les transcriptions ont été conservées. La mise à jour de Fireworks affirme que ses services d'inférence, d'entraînement et Nexus, à conservation nulle par défaut, n'ont pas été touchés. Elle range pourtant des jeux de données parmi les éléments consultés. La garantie se définit produit par produit, alors que le client achète un usage.

Ashley Madison : la suppression payée, démentie par la fuite

Le précédent le plus net date de 2015. Le 12 juillet, le site de rencontres extraconjugales Ashley Madison est piraté. En août, les profils de plus de 36 millions d'utilisateurs sont publiés. Le site vendait 19 dollars un service « Full Delete », censé effacer les données de l'utilisateur. Selon la plainte de la Federal Trade Commission, les données de ces clients avaient été conservées et figuraient parmi celles publiées. La même plainte relève que l'exploitant ne savait pas si ses prestataires appliquaient des mesures de sécurité raisonnables.

Le règlement annoncé le 14 décembre 2016 prévoit 1,6 million de dollars versés à la FTC et aux États, sur un jugement de 8,75 millions partiellement suspendu. Il impose aussi un programme de sécurité soumis à des évaluations par des tiers.

La structure est la même. Une absence de conservation est vendue comme une qualité. L'acheteur ne peut pas l'observer, et c'est une fuite qui la dément. La réponse institutionnelle consiste ensuite à ajouter un gardien, sous forme d'évaluation externe. La différence tient aux acheteurs. Les clients d'Ashley Madison étaient des particuliers. Ceux de Fireworks sont des entreprises outillées, avec contrats, questionnaires de sécurité et services juridiques. **Neuf ans plus tard, et à une autre échelle, la position de l'acheteur face à une absence promise n'a pas changé.**

L'audit annoncé ajoute un gardien de plus

Notion annonce un audit de ses fournisseurs critiques, prestataires d'inférence et sous-traitants de données, pour vérifier qu'ils respectent la minimisation des données, y compris la conservation nulle. La page ne donne ni calendrier ni méthode.

Shapiro résume le problème par une question : « So who guards the guardians? » (« Qui garde donc les gardiens ? »). L'audit répond à la défaillance d'un gardien en en installant un autre. Michael Power avait décrit en 1994, dans *The Audit Explosion*, la pente de ce type de contrôle. Il vérifie les dispositifs que l'audité déclare plutôt que l'activité elle-même. Il produit un « regulatory comfort » (un confort réglementaire) qui rassure sans garantir un examen de fond. Et il se renforce après chaque défaillance.

Appliquée à la conservation nulle, cette pente a une conséquence précise. Un audit peut constater des politiques de conservation, des paramètres de configuration, des attestations.

Il peut difficilement établir qu'aucune copie n'existe nulle part, ni sur une machine de développement ni chez un hébergeur de rang deux. Or le cas Notion ne montre pas un prestataire dépourvu de politique. Il montre une politique affichée, contredite par l'état réel des systèmes. Il montre aussi une fonction du client qui envoyait des conversations stockées hors du périmètre de sa propre promesse. **Le risque est que l'audit rassure sur la politique sans établir l'absence.**

Quand une garantie ne se constate qu'à l'occasion d'un incident

Sans la marque, le mécanisme tient en une phrase : une organisation protège une qualité qu'elle ne peut pas observer chez son prestataire par une chaîne de gardiens déclaratifs, et n'apprend la défaillance que lorsqu'un incident la rend publique.

Il se reproduit à quatre conditions. La qualité achetée doit être invisible à l'usage : conservation nulle, effacement en fin de contrat, localisation des données, engagement de ne pas entraîner de modèles sur les données du client. La protection doit reposer sur des gardiens déclaratifs (clauses, attestations, bulletins) plutôt que sur une observation possible. La chaîne de sous-traitance doit dépasser un rang. Enfin, et c'est la condition la plus discrète, le client doit lui-même créer des flux que ses propres engagements ne couvrent pas, pour des besoins d'analyse, de journalisation ou de pilotage.

> **Grille de lecture**
>
> 1. Pour chaque engagement de non-conservation signé, l'organisation dispose-t-elle de la liste des points d'accès et des fonctions qu'il couvre, ou seulement d'une formule générale ?
> 2. Le registre des traitements recense-t-il les fonctions activées (tableaux de bord d'usage, historiques, classifications) et le prestataire qui traite chacune, ou seulement les éditeurs ?
> 3. Les outils de mesure de l'usage de l'IA s'alimentent-ils du contenu des conversations, et ce contenu quitte-t-il l'éditeur principal ?
> 4. Quand l'organisation a-t-elle appris pour la dernière fois l'existence d'un sous-traitant de rang deux de l'un de ses fournisseurs : par une notification contractuelle ou par la presse ?
> 5. Quelle preuve, autre qu'une attestation du prestataire, permettrait de constater qu'une donnée n'a pas été conservée ?

Les signes sont observables de l'intérieur : des contrats qui mentionnent la conservation nulle sans en définir le périmètre, des tableaux de bord d'usage alimentés par le texte des échanges, une équipe de sécurité qui découvre la chaîne de sous-traitance d'un fournisseur au moment d'un incident.

Chronologie (à la date du 15 novembre 2026)

1. 6 octobre : Fireworks détecte l'usage non autorisé d'identifiants internes.
2. 9 octobre : premier bulletin public de Fireworks.
3. 23 et 24 octobre : compromission d'un fournisseur cloud tiers, puis des machines de développement de Fireworks.
4. 2 novembre : Fireworks signale à Notion un premier jeu de données, synthétique.
5. 3 novembre : Notion coupe ses connexions avec Fireworks.
6. 6 novembre : Fireworks signale un possible accès à des données de clients.
7. 8 novembre : confirmation du second jeu, qui contient des transcriptions Insights.
8. 9 novembre : avis de sécurité de Notion et mise à jour du bulletin de Fireworks.

Conclusion

Notion a appris que son prestataire conservait des données le jour où un groupe criminel les a emportées. En octobre, Fireworks écrivait que sa plateforme ne gardait rien. En novembre, son client écrivait le contraire. Entre les deux, la clause, le bulletin et la promesse faite aux clients Enterprise ont tenu leur rôle de gardiens sans que personne puisse observer ce qu'ils gardaient. La question de Shapiro reste ouverte : qui garde la promesse de non-conservation, quand chaque maillon ne peut que croire le suivant ?

Si cette lecture est juste, la sortie de crise ajoutera des gardiens plutôt que des moyens d'observation. Trois indicateurs permettront de le vérifier d'ici au 30 juin 2027. Notion publiera-t-il la méthode ou les résultats de son audit des fournisseurs critiques ? Insights sera-t-il réactivé avec une indication publique de l'endroit où s'exécute l'analyse des conversations ? Fireworks publiera-t-il un rapport final expliquant comment des données issues des points d'accès de Notion ont été conservées ? Un dernier indicateur est plus large : d'ici novembre 2028, un autre prestataire d'IA engagé sur la conservation nulle sera-t-il pris en défaut de la même manière ? Si c'est le cas, l'incident de novembre 2026 n'aura pas été un accident de parcours, mais le premier cas documenté d'une garantie que le marché vend sans pouvoir la montrer.

Note méthodologique

Les faits reposent principalement sur les documents de Notion et de Fireworks AI, recoupés par The Register et BleepingComputer. Les faits sont arrêtés au 15 novembre 2026. Aucune causalité n'est assumée sur l'origine de la conservation : à cette date, les deux organisations n'en décrivent pas la voie dans leurs documents publics, et l'article s'en tient à leurs déclarations. BleepingComputer écrit que moins de 1 % des utilisateurs sont touchés, alors que Notion parle de moins de 1 % des espaces de travail. Le chiffre de Notion a été retenu. L'arrêt des entraînements est attribué à BleepingComputer, car les documents de Notion

consultés ne le mentionnent pas. La confiance impersonnelle de Shapiro est le cadre central. Le *bien de croyance* de Darby et Karni, déjà mobilisé le 17 octobre 2026, n'intervient qu'en appui. Ces choix relèvent de l'auteur. Recherche de sources, croisement des données et structuration du raisonnement s'appuient sur une assistance par intelligence artificielle. Le choix de l'angle et du cadre théorique, la vérification des sources, les arbitrages et l'établissement du texte final relèvent de l'auteur.

Sources

Documents des organisations et décisions d'autorités

Notion, « Notice about the Fireworks AI incident that may impact some Notion workspaces », page d'aide, consultée le 15 novembre 2026. https://www.notion.com/help/security-incident-involving-a-third-party-ai-vendor

Notion, « Fireworks AI incident (11-09) », avis de sécurité, publié le 9 novembre 2026. https://www.notion.com/security/advisories/security-incident-involving-a-third-party-ai-vendor-nov9

Notion, « Notion AI settings overview », page d'aide, consultée le 15 novembre 2026. https://www.notion.com/help/notion-ai-settings-overview

Fireworks AI, « Fireworks October 2026 security incident », bulletin publié le 9 octobre 2026, mis à jour le 10 octobre 2026. https://fireworks.ai/security/bulletins/october-2026-security-incident

Fireworks AI, « Update on October 2026 security incident », bulletin mis à jour le 9 novembre 2026. https://fireworks.ai/security/bulletins/october-2026-security-incident-update

Presse de référence

Bill Toulas, « Notion says Fireworks breach exposed customer AI chat transcripts », BleepingComputer, 11 novembre 2026. https://www.bleepingcomputer.com/news/security/notion-says-fireworks-breach-exposed-customer-ai-chat-transcripts/

Connor Jones, « Fireworks AI gets burned twice as customer data exposure widens », The Register, 10 novembre 2026. https://www.theregister.com/cyber-crime/2026/11/10/fireworks-ai-gets-burned-twice-as-customer-data-exposure-widens/5308722

Cadres théoriques

Susan P. Shapiro, « The Social Control of Impersonal Trust », American Journal of Sociology, vol. 93, n° 3, 1er novembre 1987, p. 623 à 658. https://doi.org/10.1086/228791

Michael Power, « The Audit Explosion », Demos, Londres, 1994, version archivée le 15 janvier 2021, consultée le 15 novembre 2026. https://web.archive.org/web/20210115184145/https://

www.demos.co.uk/files/theauditexplosion.pdf

Parallèle historique

Federal Trade Commission, « Operators of AshleyMadison.com Settle FTC, State Charges Resulting From 2015 Data Breach that Exposed 36 Million Users' Profile Information », communiqué, 14 décembre 2016. https://www.ftc.gov/news-events/news/press-releases/2016/12/operators-ashleymadisoncom-settle-ftc-state-charges-resulting-2015-data-breach-exposed-36-million