RECHERCHE INSTITUTIONNELLE COMPAISS

Quand les institutions se trompent avec l'IA

Cinq cas documentés dans l'enseignement supérieur et les environnements réglementés

Mai 2026

APERÇU

Les cinq cas suivants documentent des défaillances et controverses signalées publiquement découlant du déploiement de systèmes d'IA à usage général dans des environnements institutionnels réglementés. Chaque cas est tiré de sources indépendantes largement diffusées. Ensemble, ils illustrent un schéma cohérent : les institutions qui déploient une IA axée sur l'inférence rencontrent des problèmes d'exactitude, de gouvernance et de responsabilité que les avertissements, les outils de surveillance et le filtrage après coup ne peuvent pas pleinement résoudre. Il ne s'agit pas d'incidents isolés. Ils reflètent une propriété structurelle des architectures d'IA génération-première lorsqu'elles sont appliquées dans des environnements où l'exactitude, l'autorité et la responsabilité institutionnelle sont non négociables.

CAS 01

L'Université d'État de Californie renouvelle son contrat controversé avec OpenAI

EdSource / CalMatters, mai 2026 | Kate Rix | https://edsource.org/2026/cal-state-renews-controversial-system-wide-contract-with-openai/758919

L'Université d'État de Californie a renouvelé son contrat avec OpenAI à 13 millions de dollars par année pendant trois ans, offrant un accès à l'échelle du système à ChatGPT Edu sur 22 campus et auprès de plus de 470 000 étudiants. Le renouvellement a ravivé l'opposition professorale qui s'était développée depuis la signature du contrat initial de 17 millions de dollars en janvier 2025, sans l'accord du corps enseignant.

Les professeurs ont remis une pétition formelle demandant l'annulation du contrat, affirmant que ChatGPT Edu n'est « pas conçu, formé ou optimisé pour l'éducation ». Un sondage à l'échelle du système CSU a confirmé que bien que ChatGPT soit l'outil d'IA le plus utilisé sur les campus, il a également révélé de larges préoccupations quant à l'impact futur de l'IA sur la qualité de l'enseignement, les résultats des étudiants et l'intégrité institutionnelle. Les professeurs ont exprimé des inquiétudes spécifiques concernant l'incapacité du système à citer de manière fiable des sources institutionnelles et son indifférence à l'exactitude des réponses. L'Association étudiante de l'État de Californie a rapporté que les étudiants faisaient face à des politiques de

classe incohérentes, à la crainte de fausses accusations de tricherie et à de la confusion quant aux moments où l'utilisation de l'outil sanctionné par l'université était permise.

CONSTAT CLÉ

> *Le plus grand contrat université-IA de l'histoire a généré une opposition institutionnelle soutenue précisément parce qu'un chatbot à usage général ne peut pas satisfaire aux normes de fiabilité des sources et d'exactitude que le corps enseignant et les étudiants exigent pour les orientations institutionnelles. Le problème de gouvernance n'est pas accessoire au déploiement. Il est la conséquence directe du déploiement d'une IA inférence-première dans un environnement qui requiert des réponses autorisées et spécifiques à l'institution.*

CAS 02

Les collèges communautaires de Californie dépensent des millions en chatbots IA défaillants

CalMatters, mars 2026 | https://calmatters.org/education/higher-education/college-beat/2026/03/college-ai-chatbot/

Des districts de collèges communautaires de Californie ont dépensé des millions de dollars pour déployer des chatbots IA destinés à aider les étudiants à naviguer dans les admissions, l'aide financière et les services du campus. Les reportages de CalMatters ont révélé que ces systèmes peinaient à fournir des réponses claires et exactes, laissant les étudiants frustrés et se tournant vers des canaux de médias sociaux non officiels pour obtenir de l'aide.

Certaines plateformes s'appuyaient sur des bibliothèques de FAQ maintenues manuellement et l'extraction de sites Web de campus pour générer des réponses, ce qui entraînait des erreurs lorsque les informations étaient périmées ou lorsque les questions dépassaient le cadre de formation du système. Des responsables de plusieurs districts ont reconnu les défaillances et annoncé des plans de transition vers de nouvelles plateformes. L'article documente un schéma répété dans plusieurs établissements : des collèges qui construisent ou se procurent leurs propres chatbots et rencontrent les mêmes problèmes d'exactitude et de confiance que les systèmes d'IA à usage général, parce que l'architecture inférence-première sous-jacente a été préservée.

CONSTAT CLÉ

> *Les institutions qui construisent ou se procurent leurs propres chatbots IA n'échappent pas au problème d'exactitude en s'éloignant des outils d'IA grand public. Lorsque l'architecture reste inférence-première, le mode de défaillance est le même : le système génère des réponses indépendamment du fait que des sources institutionnelles autorisées existent pour les étayer. Le contexte de déploiement change. Le risque structurel, non.*

CAS 03

L'IA devient plus puissante, mais ses hallucinations empirent

The New York Times, mai 2025 | Cade Metz et Karen Weise | https://www.nytimes.com/2025/05/05/technology/ai-hallucinations-chatgpt-google.html

Un reportage du New York Times a documenté une découverte contre-intuitive émergeant des chercheurs et développeurs en IA : à mesure que les grands modèles de langage deviennent plus

capables et largement déployés, le problème des hallucinations ne s'améliore pas au même rythme que les autres capacités. Sur certains plans, il s'aggrave.

L'article examine pourquoi des modèles plus puissants ne produisent pas automatiquement des résultats plus fiables. À mesure que les modèles sont formés pour traiter des tâches de raisonnement multi-étapes plus complexes, ils génèrent des réponses plus longues et plus élaborées, ce qui augmente la surface d'erreur factuelle. Les systèmes sont optimisés pour la fluidité et la cohérence apparente, ce qui signifie que les informations incorrectes sont présentées avec le même registre confiant que les informations correctes. Des chercheurs cités dans l'article ont décrit cela comme une tension fondamentale dans la façon dont les grands modèles de langage sont formés et évalués.

CONSTAT CLÉ

> *Les améliorations de capacité dans les modèles d'IA généraux ne résolvent pas le problème d'exactitude dans les environnements institutionnels réglementés. Un modèle plus capable qui produit des réponses plus longues, plus élaborées et énoncées avec plus d'assurance introduit davantage de façons pour la désinformation institutionnelle d'être prise en compte avant d'être identifiée et corrigée. La solution n'est pas un modèle plus puissant. C'est une approche architecturale différente.*

CAS 04

Stanford HAI et MIT Sloan : taux d'hallucination dans les applications IA à enjeux élevés

MIT Sloan Teaching and Learning Technologies, citant Stanford HAI et les recherches d'OpenAI, 2025 | https://mitsloanedtech.mit.edu/ai/basics/addressing-ai-hallucinations-and-bias/

Une évaluation par les pairs compilée par MIT Sloan, s'appuyant sur les recherches de Stanford HAI et les propres conclusions publiées d'OpenAI, a documenté les taux d'hallucination empiriques des principaux systèmes d'IA dans des domaines de requêtes à enjeux élevés. Les résultats comptent parmi les plus cités dans la littérature sur la gouvernance institutionnelle de l'IA.

L'étude de Stanford HAI a constaté que les chatbots IA à usage général ont halluciné sur 58 à 82 pour cent des requêtes de recherche juridique lors des tests sur des modèles contemporains. Même les outils juridiques IA spécialisés construits sur la génération augmentée par récupération (RAG), une technique qui ancre les réponses IA dans une base de données documentaire triée, ont halluciné plus de 17 pour cent du temps. Séparément, les propres recherches d'OpenAI ont révélé que les modèles de raisonnement avancés n'éliminaient toujours pas systématiquement les hallucinations lors du raisonnement multi-étapes. Les modèles font des « conjectures stratégiques », générant des énoncés plausibles mais faux lorsqu'ils sont incertains, et sont dans certains cas involontairement récompensés pour avoir halluciné lors de la formation et de l'évaluation.

CONSTAT CLÉ

> *Même les contrôles post-génération les plus sophistiqués, y compris la génération augmentée par récupération, ne réduisent pas les taux d'hallucination à zéro dans les domaines réglementés. À 17 pour cent d'hallucination résiduelle dans des conditions RAG optimales, et 58 à 82 pour cent sans*

> *ancrage spécialisé, les taux d'erreur dans les contextes institutionnels à enjeux élevés ne sont pas des cas limites d'ingénierie. Ce sont des conséquences documentées, mesurées et structurellement prévisibles des architectures inférence-première.*

CAS 05

CSU et OpenAI : quand les étudiants refusent l'IA que leur université leur impose

CalMatters, mai 2026 | https://calmatters.org/education/2026/05/california-state-university-open-ai-chatgpt-contract/

Un article complémentaire sur le renouvellement du contrat a documenté les conséquences humaines du déploiement d'IA à usage général à l'échelle institutionnelle sans architecture de gouvernance alignée sur les obligations institutionnelles. Des étudiants de plusieurs campus CSU ont rapporté avoir été accusés de tricherie pour avoir utilisé l'outil IA que le système universitaire avait mandaté et promu. Le personnel des centres d'aide aux devoirs a rapporté être incapable de conseiller les étudiants pris entre des politiques professorales contradictoires sur l'utilisation de l'IA.

L'Association étudiante de l'État de Californie a constaté que les étudiants vivaient de la « confusion, de la peur et de la méfiance » comme conséquence directe de l'absence d'orientations cohérentes et spécifiques à l'institution sur ce que l'IA pouvait et ne pouvait pas être utilisée pour fournir. 83,5 pour cent des étudiants dans un sondage à l'échelle du système ont rapporté des préoccupations concernant l'impact de l'IA sur leurs données personnelles. Des professeurs ont cité une exposition juridique potentielle découlant du déploiement d'un système qui avait été lié, dans des dépôts judiciaires en Californie, à des allégations de préjudice psychologique. L'article documente un vide de gouvernance créé lorsqu'un outil à usage général puissant est déployé sans les contrôles d'autorité institutionnelle que les environnements réglementés exigent.

CONSTAT CLÉ

> *Les conséquences du déploiement d'IA inférence-première à l'échelle institutionnelle ne sont pas seulement techniques. Lorsque les étudiants ne peuvent pas faire confiance qu'un outil IA endossé par leur établissement leur fournira des informations exactes et autorisées sur les politiques institutionnelles, les échéances et les droits, le résultat n'est pas simplement de la désinformation. C'est une érosion de la confiance institutionnelle que les organisations réglementées ont l'obligation de maintenir.*

Synthèse : le schéma commun à tous les cinq cas

Ces cinq cas couvrent différents types d'institutions, différents fournisseurs d'IA, différents contextes de déploiement et différents organes de presse. Ils partagent un constat structurel commun.

Dans chaque cas, l'institution a déployé ou envisagé de déployer un système d'IA construit sur une architecture inférence-première : un système conçu pour générer une réponse par défaut, avec un risque géré par des contrôles post-génération, des avertissements, une surveillance ou

une supervision humaine. Dans chaque cas, cette architecture a produit des défaillances en matière d'exactitude, de responsabilité ou de gouvernance que les contrôles post-génération ne pouvaient pas pleinement prévenir.

Les défaillances ne sont pas des défaillances de produits ou de fournisseurs individuels. Ce sont des conséquences prévisibles d'une hypothèse architecturale incompatible avec les obligations des environnements institutionnels réglementés : l'hypothèse que l'inférence devrait toujours se produire, et que les risques de la génération toujours active peuvent être gérés après coup.

COMPAiSS répond à cela au niveau architectural en conditionnant l'existence de l'inférence à une autorisation pré-exécution contre des sources approuvées par l'institution. Lorsqu'aucune base d'autorité n'existe pour une réponse, le modèle ne s'exécute pas. Il n'y a pas de réponse à filtrer, pas d'hallucination à attraper, pas de désinformation institutionnelle à corriger après qu'un étudiant ait déjà agi sur celle-ci.
