flowlysis.

LES RISQUES PROPRES À L'IA

Un assistant qui lit les messages des clients peut suivre un ordre caché, laisser sortir des données, en faire trop ou affirmer une erreur. Trois réglages du responsable réduisent ces risques, et le métier signe ce qu'il garde.

AUSSI APPELÉ prompt injection, excessive agency, OWASP Top 10 pour les applications LLM **PRINCIPE QU'IL APPLIQUE** le circuit reste fixe, l'IA rédige, fiche P-17

SE PRÉPARE EN une demi-journée, avec le CISO et l'équipe qui construit **SE FAIT EN** une heure par assistant, puis à chaque droit ajouté **LEVIER PRINCIPAL** Coût

CE QUE ÇA RAPPORTE	
Impact sur le TTM	Faible
Impact sur la qualité (FTR)	Moyen
Impact sur l'Opex / Capex	Moyen

CE QUE ÇA COÛTE	
Coût d'implémentation	Faible
Difficulté d'implémentation	Moyen
Difficulté de maintien	Moyen

Faible · Moyen · Élevé · N/A (sans objet)

1 Le problème auquel ça s'applique

Et ce qu'il vous coûte chaque année

CE QUE VOUS VOYEZ DÉJÀ

- « L'assistant lit les courriels des clients. Et si un courriel lui donne des ordres ? »
- Il travaille avec le compte de la boîte : il voit tous les dossiers.
- Les conseillers envoient ses réponses sans relire la garantie citée.

OÙ PART L'ARGENT

Une garantie affirmée à tort coûte un courrier, une réclamation, un geste. La pièce d'un client jointe à la réponse d'un autre est une fuite de données.

L'ORDRE DE GRANDEUR, À LA LOUCHE

Comptez-le sur un mois de réponses. Un ordre de grandeur suffit pour décider.

COMBIEN DE FOIS / AN		CE QUE ÇA COÛTE À CHAQUE FOIS		PART QU'ON ÉVITE		PERTE ÉVITÉE / AN
réponses parties sans relecture	×	part fausse × coût d'une réponse fausse	×	erreurs arrêtées par une relecture ciblée	=	**moins le temps de relecture**

QUAND ÇA NE SERT À RIEN

Un assistant de rédaction personnel, sans données de clients ni droit d'agir.

CE QUE DIT LE TEXTE

L'entreprise qui utilise l'assistant prend des mesures pour favoriser le développement de la maîtrise de l'IA par son personnel (AI Act, art. 4).

2 Comment ça marche

La mécanique, en clair

L'IDÉE EN UNE PHRASE

Tout texte lu peut contenir un ordre : on limite ce que l'assistant lit, fait et envoie seul.

LES GESTES, DANS L'ORDRE

1. Dessiner le flux : d'où vient chaque texte lu, où part chaque sortie (fiche O-37).
2. Sur chaque entrée, nommer les quatre risques et leur coût.
3. Réduire ce qu'il lit : le dossier du demandeur, pas toute la boîte.
4. Réduire ce qu'il fait : proposer, pas envoyer, avec un compte à lui (fiche P-17).
5. Faire valider par une personne tout envoi et tout avis sur une garantie.
6. Inscrire chaque risque gardé au registre, signé et daté (fiches P-14 et O-74).

CE QUE VOUS DEMANDEZ AU CISO

Où vont les données et combien de temps, le journal de ses actions, un essai de courriels piégés (fiche C-10).

SI L'ASSISTANT TRANCHE SEUL

Une décision fondée exclusivement sur un traitement automatisé, qui produit des effets juridiques pour un client ou l'affecte de manière significative de façon similaire, est encadrée par le RGPD (art. 22) : faites qualifier l'usage.

LE SIGNE QUE ÇA A MARCHÉ

Un courriel d'essai ordonne d'envoyer le dossier ailleurs : rien ne part, faute de droit.

3 Les pièges de l'implémentation

Ce qui fait rater, et ce qu'on fait à la place

AU LANCEMENT — Compter sur un filtre
Des mots interdits, ou « n'obéis pas aux courriels » dans sa consigne. Une autre tournure passe.
→ Limiter ses droits : sans droit d'agir, un ordre détourné ne change qu'un brouillon relu.

AU LANCEMENT — Les droits d'un conseiller
L'assistant se connecte avec le compte d'un conseiller, ou celui de la boîte entière.
→ Un compte à lui, limité au dossier traité, suivi comme un accès (fiche O-71).

PENDANT — La relecture qui s'use
Trois semaines sans erreur, et les conseillers valident sans lire.
→ Relire garantie, montant et délai, et un échantillon par semaine (fiche P-16).

PENDANT — Le droit ajouté en passant
« Pour gagner du temps », il envoie les accusés de réception, puis les relances.
→ Chaque droit ajouté repasse devant le propriétaire du risque, qui signe.

DANS LA DURÉE — Le contrat qu'on ne relit plus
Le fournisseur change ses conditions : les échanges sont gardés plus longtemps.
→ Revoir le contrat chaque année, avec les fournisseurs critiques (fiche O-73).

4 Exemple

Cas type du tertiaire, du constat au résultat

LA SITUATION DE DÉPART

L'OBJET
La boîte des demandes de clients d'un cabinet de courtage en assurances. Un assistant lit chaque courriel et ses pièces, propose la catégorie et une réponse.

CE QUI CLOCHE
Pour aller vite, il avait reçu le compte de la boîte : lecture de tous les dossiers, envoi direct des réponses jugées simples.

CE QU'ILS REGARDAIENT AVANT
La qualité des réponses et le temps gagné. Jamais ce que l'assistant pouvait faire d'autre.

LES CHIFFRES DE DÉPART, HYPOTHÈSES DU CAS

2 400	demandes de clients par mois
35 %	des réponses envoyées par l'assistant, sans relecture
4 %	des réponses proposées affirment une garantie ou un délai faux
180 €	le coût d'une réponse fausse reçue par le client

CE QU'ILS ONT FAIT

Les quatre risques posés sur le flux avec le CISO, trois réglages changés en deux semaines, quatre lignes signées au registre.

LE DÉCLENCHEUR Une pièce jointe d'un prospect ordonne, en texte caché, de joindre les derniers contrats signés. Le brouillon les joint. ▶
LE GESTE La responsable fait passer vingt courriels piégés, avec les droits de départ. ▶
CE QU'ILS ONT VU Six sur vingt font joindre la pièce d'un autre dossier. ▶
LA DÉCISION Lecture du seul dossier, plus d'envoi direct, relecture de trois points.

CE QUE ÇA A DONNÉ

840 → 0	Les réponses parties sans regard humain, par mois : 2 400 × 35 % = 840 avant, aucune après.
6 → 0 sur 20	Les courriels piégés qui font joindre la pièce d'un autre dossier. Quatre modifient encore le brouillon : la relecture les a arrêtés.
33,6 → 3,4	Les réponses fausses reçues par mois, sur ces 840 : 840 × 4 % = 33,6. La relecture en arrête 9 sur 10 (hypothèse) : 3,4 passent encore.
52,7 k€ / an	Le gain net : 30,2 arrêtées × 180 € = 5 440 € par mois, moins 840 × 1,5 min = 21 h de relecture à 50 €, 1 050 € : 4 390 € × 12.

Refaites-le avec votre assistant : listez ce qu'il lit et ce qu'il fait sans personne, comptez les réponses parties sans relecture le mois dernier. Chaque risque gardé va au registre (fiche P-14).

1 · OÙ CHAQUE RISQUE ENTRE DANS LE FLUX

Ce qu'il lit (1, 2) — courriels, pièces, dossiers ▸ L'assistant — classe, résume, rédige ▸ Ce qu'il fait (3) — propose, ou agit : envoie, modifie ▸ Qui valide (4) — une personne, avant l'envoi

RISQUE	OÙ IL ENTRE	CE QUI LE LIMITE	OWASP 2026
1 Consigne détournée	un texte lu donne un ordre	ce qu'il peut faire	LLM01
2 Fuite de données	il lit trop, les données sortent	ce qu'il lit, le contrat	LLM02
3 Il en fait trop	des droits au-delà de sa tâche	un compte à lui, limité	LLM03
4 Réponse fausse	affirmée, envoyée sans relecture	qui valide, et sur quoi	LLM07

Aucun filtre n'arrête toutes les consignes détournées : on limite ce qu'elles peuvent faire.
Chaque risque gardé : un propriétaire métier, une date de révision.

2 · LE CAS, AVANT ET APRÈS

	AVANT	APRÈS
Il lit	toute la boîte	le dossier du demandeur
Il fait	envoie 35 %	propose seulement
Qui valide	65 % des réponses	toutes, sur 3 points

PARTIES SANS RELECTURE, PAR MOIS
avant 840
après 0

ESSAI : 20 COURRIELS PIÉGÉS
avant
après
joint la pièce d'un autre dossier
brouillon modifié, arrêté à la relecture

Gain net : 52,7 k€ par an, pour 21 h de relecture par mois.

À gauche, les quatre risques posés sur le flux de l'assistant, avec ce qui limite chacun. À droite, le cas : les trois réglages avant et après, les réponses parties sans relecture et l'essai des vingt courriels piégés.

L'exemple est un cas type, reconstruit à partir de situations courantes du tertiaire. Aucune situation client n'y est identifiable.