COLLABOR-IA · FICHE-MÉMO

Ce que l'assistant fait seul, ce qu'il vous demande

Ce que la règle apporte.
Chacun sait d'avance ce que l'assistant mène seul, ce qui attend un « oui », et de qui.

Ce qu'elle évite.
Un mail parti chez le mauvais client, un virement lancé sur la foi d'un message piégé, un dossier effacé sans que personne le voie partir.

Ce que contient la fiche.
Une échelle de cinq actions, le valideur de chacune, les six éléments d'une demande de confirmation, la trace à garder, puis huit questions.

Pour les dirigeants et responsables d'équipe de TPE et PME dont l'assistant IA peut lire, écrire ou envoyer

Édition du 19/11/2026 · NLQ

01 Rangez chaque action sur l'échelle

Un assistant branché à la messagerie ou au logiciel de gestion ne fait pas qu'une chose. Il lit, rédige, envoie, parfois règle ou efface. L'ANSSI recommande qu'aucune action critique ne soit exécutée sans contrôle humain : elle cite les transactions bancaires, la production de contenu public et ce qui touche directement des personnes (R9).

Lire	Préparer	Envoyer	Payer	Supprimer
chercher, résumer	brouillon, proposition	mail, devis, publication	virement, commande	fichier, fiche, message

← MOINS DE RISQUE PLUS DE RISQUE →

Des verbes, pas des outils.
« Répondre à une demande de devis », « relancer une facture », « archiver un dossier clos » : une ligne par action réelle de l'équipe.

La conséquence fait la marche.
On classe selon ce qui arrive si l'action est fausse, pas selon sa difficulté. Une relance mal adressée est un envoi, même si elle tient en deux lignes.

Dans le doute, la marche au-dessus.
Une action qu'on ne sait pas classer prend la marche la plus haute des deux.

Ce qui devient public est un envoi.
Une publication sur le site ou un réseau social est du contenu public au sens de l'ANSSI : jamais automatique.

02 Laissez faire seul ce qui ne sort pas

Lire et préparer ne changent rien et ne sortent rien de l'entreprise. C'est la frontière que décrit l'OWASP : un assistant qui résume les mails peut rédiger des réponses, mais c'est l'utilisateur qui relit et appuie sur « Envoyer », chaque fois.

Le brouillon reste un brouillon.
L'assistant prépare dans le dossier des brouillons, jamais dans la boîte d'envoi.

Rien ne part tout seul après une lecture extérieure.
Quand l'assistant vient de lire un mail reçu ou une page web, l'action suivante attend un humain (ANSSI R27).

Le gain reste entier.
Recherche, tri et rédaction sont faits : il ne reste à la personne que relire et décider.

03 Nommez qui valide chaque marche

Une validation n'a de valeur que si la personne qui dit « oui » sait juger. Le plus simple est de garder le circuit qui existait avant l'assistant.

MARCHE	QUI VALIDE	EXEMPLE FICTIF
Lire, préparer	Personne : l'assistant fait seul	Résumer les demandes du jour
Envoyer	La personne qui signe le message	La chargée d'accueil envoie la réponse
Payer	La personne qui signe déjà les virements	Le gérant valide le règlement fournisseur
Supprimer	Le responsable du dossier, sauvegarde vérifiée	Le chef d'équipe archive puis efface

Le même circuit qu'avant.
Celui qui signait un virement signe celui que prépare l'assistant. Rien ne change dans la chaîne de décision.

Un suppléant par marche.
Sinon, l'absence d'une personne bloque l'équipe, ou pousse à tout débloquer.

Un changement de coordonnées bancaires se vérifie ailleurs.
Un message qui demande de payer sur un nouveau compte se confirme auprès de l'émetteur par un autre moyen, comme le recommande Cybermalveillance.gouv.fr.

La règle vit dans l'outil.
L'OWASP recommande que l'autorisation soit appliquée par le système, pas laissée au jugement du modèle : l'assistant n'a pas le droit de payer, il prépare.

04 Exigez une demande de confirmation qui montre tout

Pour la spécification MCP, l'application doit signaler chaque appel d'un outil, permettre de le refuser et afficher ce qui sera envoyé avant l'appel, pour éviter qu'une donnée parte par erreur ou par malveillance. Une demande de confirmation utile ressemble à ceci :

```
L'assistant demande à ENVOYER un mail
Outil          messagerie de l'entreprise
À              compta@client-fictif.example.com
Objet          Relance facture F-2026-118
Contenu        texte complet ci-dessous, pas un résumé
À l'origine    mail reçu ce matin, « Service facturation »
[ Refuser ]   [ Modifier ]   [ Envoyer ]
```

L'action en un verbe.
Envoyer, payer, supprimer : écrit en tête, sans détour.

Le destinataire ou le montant en clair.
L'adresse complète, le compte crédité, le fichier visé.

Le contenu entier.
On valide le texte qui part, pas le résumé qu'en fait l'assistant.

L'origine de la demande.
Quel mail ou quel document a déclenché l'action. Une demande née d'un message extérieur se relit deux fois.

05 Gardez la trace des oui et des non

L'ANSSI recommande de journaliser les traitements de l'assistant, y compris ses appels à d'autres outils, pour pouvoir reconstituer un événement (R29). La spécification MCP demande aussi de garder la trace de l'usage des outils.

Qui, quoi, quand.
Pour chaque validation : la personne, l'action, la date, la réponse.

Les refus d'abord.
Une fois par mois, relisez ce qui a été refusé : c'est là que l'on voit ce que l'assistant tente de faire de travers.

Une erreur fait monter la marche.
Une action validée par habitude et qui s'est révélée fausse remonte d'un cran sur l'échelle.

La sauvegarde avant l'effacement.
La sauvegarde est souvent le seul moyen de retrouver des données (Cybermalveillance.gouv.fr). On vérifie qu'elle existe avant d'autoriser la marche « Supprimer ».

Comment savoir si c'est en place

Huit questions fermées. Chaque « non » renvoie à une règle de la fiche.

- [ ] Existe-t-il une liste écrite des actions que l'assistant peut faire?
- [ ] L'assistant est-il incapable d'envoyer sans un clic humain?
- [ ] L'assistant est-il privé du droit de payer par l'outil lui-même?
- [ ] Les refus du mois dernier ont-ils été relus?
- [ ] Chaque action est-elle rangée sur une marche?
- [ ] Chaque marche a-t-elle un valideur et un suppléant nommés?
- [ ] La demande de confirmation montre-t-elle le contenu entier et l'origine?
- [ ] La sauvegarde est-elle vérifiée avant toute suppression?

VOTRE PROCHAINE ÉTAPE

Apportez la liste des actions que votre assistant peut faire: on la range ensemble sur l'échelle et on écrit qui valide quoi.

Réserver un échange gratuit

www.collabor-ia.fr/rendez-vous/

Sources et portée

Toutes les sources ont été ouvertes le 19/11/2026.

- ANSSI, guide ANSSI-PA-102 *Recommandations de sécurité pour un système d'IA générative*, version 1.0 du 29/04/2024: R9, R27, R29. Recommandations sans caractère normatif. messervices.cyber.gouv.fr
- Spécification MCP 2026-07-28, *Tools*: humain dans la boucle, indicateurs d'appel, affichage des entrées, journal d'usage. modelcontextprotocol.io/specification/2026-07-28/server/tools
- OWASP Gen AI Security Project, *LLM06:2025 Excessive Agency*: validation humaine des actions à fort impact, exemple du mail relu puis envoyé par l'utilisateur, autorisation dans les systèmes en aval. genai.owasp.org/llmrisk/llm062025-excessive-agency/
- Cybermalveillance.gouv.fr, *Les 10 mesures essentielles pour votre cybersécurité*: sauvegarde, confirmation par un autre moyen. cybermalveillance.gouv.fr

Portée: les exemples, les personnes et les adresses sont fictifs. Cette fiche ne traite ni des délégations de pouvoir ni des règles comptables de l'entreprise, qui restent celles de l'entreprise.

Édité par NLQ — Collabor-IA (www.collabor-ia.fr), avec le soutien d'ALMA, sponsor. Édition du 19/11/2026. Exemples fictifs ; ce qui est dit des textes cités vient des sources datées ci-dessus.