ÉQUIPES ET CONTRÔLE

IA agentique : définition, fonctionnement et contrôle des agents en PME



Par Pierre-Louis Gaudreau, Président d'IA.consulting · Mis à jour le 20 octobre 2026

Un agent IA, qu'est-ce que c'est, et qui voit l'erreur avant votre client ?

Le cas se présente souvent ainsi. Un éditeur vous propose un « agent » qui traite les demandes clients, relance les impayés ou prépare les commandes. Il travaille seul, dit la plaquette. Vous voulez savoir deux choses. Qu'est-ce que c'est, exactement ? Et s'il se trompe, qui s'en aperçoit avant votre client ?

Définition de l'IA agentique

On appelle IA agentique des systèmes d'IA capables de poursuivre un objectif en plusieurs étapes, en choisissant eux-mêmes l'action suivante. Un agent IA reçoit un but (« traiter cette réclamation »). Il consulte des informations, choisit ses outils (votre messagerie, votre logiciel de gestion, un tableur), agit, vérifie le résultat et recommence si besoin.

Trois traits le séparent d'un logiciel ordinaire :

- **il agit** sur vos systèmes, au lieu de produire seulement un texte ;
- **il enchaîne** plusieurs étapes sans qu'une personne valide chacune ;
- **il choisit** son chemin, ce qui rend son comportement moins prévisible qu'un programme écrit à l'avance.

Quand plusieurs de ces systèmes se partagent le travail (l'un trie, l'autre rédige, un troisième contrôle), on parle de systèmes multiagents.

IA agentique et IA générative : la différence qui compte

L'IA générative produit un contenu à la demande : un texte, une synthèse, une image. Vous lisez, vous décidez, vous agissez. L'IA agentique s'appuie souvent sur un modèle génératif pour raisonner, et elle ajoute la main. Elle envoie le message, modifie la fiche client, déclenche la commande.

Pour vous, la différence tient en un mot : responsabilité. Avec l'IA générative, l'erreur passe par une personne avant d'atteindre l'extérieur. Avec un agent autonome, elle peut sortir

directement. Tout le reste de cette page découle de là. La définition détaillée de l'IA générative se trouve dans notre page sur l'impact de l'IA sur les modèles économiques (A).

Ce que disent les chiffres en France et aux États-Unis

En France : l'IA reste surtout l'affaire des grandes entreprises

En 2025, 18 % des entreprises françaises de 10 salariés ou plus utilisent au moins une technologie d'IA [S1]. L'écart selon la taille est net : 58 % des entreprises de 250 salariés ou plus, contre 15 % de celles de 10 à 49 salariés [S1].

Une précision change la lecture. L'Insee mesure l'usage de l'IA au sens large, sans isoler les agents autonomes [S1]. Ces chiffres ne disent donc ni combien de PME en utilisent, ni si ces outils sont fiables. Ils montrent seulement que la plupart des petites entreprises partent de loin. Elles aborderont ces systèmes sans équipe dédiée à leur surveillance.

Côté règles, la CNIL et le Conseil de l'IA et du numérique ont publié en juillet 2026 une analyse des systèmes agentiques [S3]. Nous en retenons trois repères :

- **traçabilité** : on doit pouvoir reconstituer tout le cheminement de décision du système [S3] ;
- **minimisation** : seules les données strictement nécessaires à la tâche doivent être traitées [S4] ;
- **sécurité d'ensemble** : une faille dans un seul composant peut servir à atteindre tout le système [S5].

Ce sont des recommandations de principe. Elles ne fixent ni format ni durée de conservation des journaux [S3]. Le constat sur la sécurité est qualitatif : aucune mesure de fréquence ni de dégâts [S5].

Aux États-Unis : le détournement, un problème qui va durer

Le NIST est l'organisme fédéral américain des normes. Par son Center for AI Standards and Innovation, il considère que le détournement d'agents (en anglais, agent hijacking) restera un problème tant que ces systèmes évolueront [S6]. Le principe : glisser des instructions dans un contenu que le système lit (un message, une page web, un document) pour lui faire exécuter autre chose que sa tâche. Le NIST en tire une conséquence pratique. Il faut des attaques simulées, menées par des testeurs (le red teaming), et les répéter en continu [S6].

Lors d'une compétition d'attaques simulées à grande échelle, au moins une attaque a réussi contre chacun des modèles les plus avancés testés [S7]. Gardez la limite en tête. Ce sont des conditions de compétition, qui ne donnent ni le taux de réussite en usage réel ni l'efficacité de telle ou telle parade [S7]. Le billet du NIST sur le détournement porte lui aussi sur des essais en laboratoire, et ne dit rien de la fréquence des incidents en entreprise [S6].

Ce que nous en tirons : aucun modèle n'est à l'abri. La protection tient donc à ce que vous autorisez le système à faire, plus qu'au modèle lui-même.

Comment fonctionne un agent IA : percevoir, raisonner, agir, apprendre

Les quatre temps du cycle

1. **Perception.** Le système reçoit des entrées : un message, un formulaire, une ligne de tableur, une alerte. C'est aussi par là qu'entrent les détournements décrits plus haut.
2. **Raisonnement.** Le modèle interprète la demande, découpe l'objectif en étapes et choisit l'outil.
3. **Action.** Il appelle l'outil : il écrit dans le fichier clients, envoie un message, crée un bon de commande. C'est à ce moment que l'erreur devient visible de l'extérieur.
4. **Apprentissage.** Il garde en mémoire ses échanges ou vos corrections, et ajuste ses choix suivants. Dans les outils courants, cette mémoire est en général un historique de conversation ; le modèle lui-même n'est pas réentraîné.

Le risque se concentre aux deux bouts. Ce que le système lit. Et ce qu'il a le droit de faire.

Exemples d'IA agentique par fonction

Fonction	Exemple de tâche confiée	Ce qui peut mal tourner
Service client	Lire une réclamation, retrouver la commande, proposer un geste commercial	Geste accordé hors politique, réponse envoyée au mauvais client
Administration des ventes	Saisir dans le logiciel de gestion une commande reçue par courriel	Mauvaise référence, quantité mal lue
Comptabilité	Rapprocher factures et paiements, préparer les relances	Relance envoyée à un client qui a déjà payé
Achats	Comparer des devis fournisseurs, préparer la commande	Fournisseur retenu sur un critère erroné
RH	Trier des candidatures, fixer des entretiens	Données personnelles traitées au-delà du besoin

Votre besoin relève-t-il d'un agent, d'une automatisation classique ou d'un assistant qui se contente de répondre ? Notre page E3 compare les trois.

Ce qu'une PME peut y gagner

- **Moins de ressaisie** entre des logiciels qui ne se parlent pas.
- **Les demandes courantes traitées hors horaires**, le reste repris par l'équipe le lendemain.
- **Des tâches à plusieurs étapes prises en charge**, là où une automatisation classique s'arrête dès qu'un cas sort du modèle prévu.

Ces gains supposent un cadre. Leur chiffrage en euros est traité dans notre page C4 sur le retour sur investissement de l'IA.

Ce qui peut mal tourner

- **Erreurs silencieuses** : le système agit avec assurance, même quand il se trompe.
- **Détournement** : des instructions cachées dans un contenu qu'il lit [S6].
- **Effet domino** : une faille dans un composant peut compromettre tout le système [S5].
- **Opacité** : sans journal, impossible de comprendre après coup pourquoi il a agi ainsi [S3].
- **Données** : il accède souvent à plus d'informations qu'il n'en a besoin [S4].

Le contrôle en cinq points

Les pages d'éditeurs énumèrent ces risques. Reste à savoir ce que vous pouvez exiger avant de le laisser agir seul.

1. Tester avant la mise en service. Faites-lui traiter des cas réels passés, sans action réelle : il propose, une personne compare avec ce qui avait été fait. Ajoutez des cas pièges : demande ambiguë, client inconnu, message contenant une consigne du type « ignore tes instructions ». Le NIST recommande d'en faire une habitude, à reprendre régulièrement, plutôt qu'un contrôle unique [S6].

2. Limiter les droits d'accès. La CNIL pose que seules les données strictement nécessaires à la tâche doivent être traitées [S4]. Le texte vise les données personnelles. L'étendre aux droits techniques (quels outils, quelles actions) est une lecture prudente que nous proposons, et non une exigence écrite [S4]. En pratique : lecture seule partout où l'écriture n'est pas indispensable, aucun accès aux paiements, plafond sur les gestes commerciaux.

3. Faire valider les actions qui engagent. Rangez les actions en deux familles. Celles qui restent internes et se défont sans dommage (brouillon, classement, note) peuvent être automatiques. Celles qui engagent l'entreprise vers l'extérieur (envoi au client, commande, remboursement) passent par une validation humaine, au moins pendant le rodage.

4. Tenir un journal. Chaque décision doit pouvoir être reconstituée : ce que le système a lu, ce qu'il a décidé, quel outil il a appelé, avec quel résultat [S3]. La CNIL ne fixe ni format ni durée [S3]. À vous de choisir. Un journal que personne ne sait lire ne sert à rien.

5. Prévoir la reprise en main. Nommez la personne qui reprend le dossier, le canal par lequel elle est prévenue et le délai de réponse attendu. Prévoyez aussi l'arrêt : qui peut couper le système, et comment.

Le passage à l'an 2000 sert de comparaison, pas de preuve. Le risque avait été très préparé, jusqu'à un courrier envoyé aux patrons de PME en décembre 1998. Le passage s'est fait sans perturbation majeure, ce qui ne prouve pas que la préparation était inutile. La comparaison s'arrête à l'échéance, unique en 2000, alors qu'un agent peut se tromper n'importe quel jour : les essais sont à reprendre.

Grille de mise en service

Cochez chaque ligne avant d'ouvrir l'outil à vos clients ou à vos fournisseurs.

Point de contrôle	Question à poser	Feu vert si
Essais	A-t-il traité des cas passés et des cas pièges sans action réelle ?	Les écarts sont compris et corrigés
Droits d'accès	Chaque accès est-il justifié par la tâche ?	Aucun droit d'écriture ou de paiement en trop
Validation	Les actions vers l'extérieur sont-elles validées par une personne ?	Oui, au moins pendant le rodage
Journal	Peut-on reconstituer une décision de bout en bout ?	Une personne non technicienne sait le lire
Reprise	Qui reprend la main, par quel canal, sous quel délai ?	Un nom, un canal, un délai, par écrit
Arrêt	Qui peut couper le système ?	Au moins deux personnes, procédure essayée
Relecture	Qui relit un échantillon d'actions chaque semaine ?	Un responsable désigné

Cas construit, pour illustrer : une PME de négoce de fournitures industrielles confie à un agent la saisie des commandes reçues par courriel. Pendant le rodage, il prépare la

commande en brouillon dans le logiciel de gestion ; l'assistante commerciale la valide d'un clic. Les commandes hors catalogue, ou d'un client inconnu, lui sont envoyées directement. Chaque vendredi, la responsable de l'administration des ventes relit un échantillon de commandes et la liste des cas renvoyés. Le système ne voit ni les conditions tarifaires négociées ni les coordonnées bancaires. Ce cadre n'empêche pas l'erreur. Il garantit qu'une personne la voit avant le client.

Ce que les pages d'éditeurs ne disent pas

Qui voit l'erreur avant le client

Les éditeurs décrivent ce que leur outil sait faire. Ils posent rarement la question inverse : à quel endroit de la chaîne une erreur est-elle repérée ? Dans une PME, la réponse est souvent « nulle part », faute de quelqu'un de désigné. Une règle : pour chaque action qui sort de l'entreprise, nommez la personne qui la voit en premier. Si c'est le client, le cadre n'est pas terminé.

Que se passe-t-il quand l'outil est bloqué

Un système bloqué qui le dit, cela se gère. Le cas dangereux est celui qui ne sait pas qu'il est bloqué et invente une réponse. Exigez de votre prestataire ou de votre outil trois comportements. Signaler qu'il n'a pas pu conclure, transmettre le dossier à une personne avec ce qui a déjà été fait, ne jamais agir « par défaut ». Vérifiez-le vous-même avec une demande impossible.

Le cadre CNIL, lu comme une liste de contrôle

L'analyse CNIL / CIANum apparaît dans les premiers résultats de recherche, mais peu de pages s'en servent. Lue par un dirigeant, elle se traduit ainsi : traçabilité, un journal lisible [S3] ; minimisation, des droits réduits à la tâche [S4] ; risque d'ensemble, des outils cloisonnés pour qu'une faille ne contamine pas tout [S5]. Le texte n'a pas la force d'une norme technique. Il donne pourtant la grille qu'une autorité française applique déjà à ces systèmes, et des arguments face à un éditeur pressé.

Décider : quand confier une tâche à un agent, et les limites

Ce type d'outil a sa place quand la tâche comporte plusieurs étapes, revient souvent, et qu'une erreur peut encore être rattrapée avant d'atteindre l'extérieur. Il en a moins quand chaque erreur coûte cher et se voit tout de suite (paiement, engagement contractuel), ou quand le volume est trop faible pour payer le temps de surveillance.

Les limites de cette page sont nettes. Les chiffres disponibles mesurent l'usage de l'IA en général [S1] et des résultats d'essais en laboratoire [S6] [S7]. Ils ne disent pas à quelle fréquence un agent se trompe dans une PME. Le contrôle décrit plus haut réduit le risque

sans le supprimer. Pour un outil qui se borne à lire et à préparer des brouillons, le risque est faible et un contrôle léger suffit. Pour un outil qui écrit chez vos clients, ce n'est plus le cas.

D'autres pages du dossier traitent les sujets voisins : démarrer l'automatisation (page D), former les équipes à surveiller ces outils (page F1), choisir le prestataire qui l'installera (page E2). Les listes d'outils agentiques gratuits sortent du périmètre de ce site. Le cadre d'ensemble est dans le guide complet pour installer l'IA dans l'entreprise (ia.consulting/fr/ia-pme/equipes-controle/). Pour savoir qui fait tourner et surveille un modèle dans la durée, lisez notre fiche sur le MLOps (ia.consulting/fr/metiers-ia/mlops/).

Et ensuite : qui porte la surveillance

Une fois l'outil défini et le contrôle posé, il reste une question d'organisation. Qui, dans l'équipe, assure la surveillance, avec quelles données et quelles règles internes ? C'est l'objet de la page principale du dossier, qui détaille ce qu'il faut organiser côté équipes, données et contrôle pour installer l'IA (ia.consulting/fr/ia-pme/equipes-controle/), au-delà du cas des agents.

Sources

- **S1** · Insee, Insee Première n° 2120, juillet 2026. Usage de l'IA dans les entreprises. https://www.insee.fr/fr/statistiques/9025878 (https://www.insee.fr/fr/statistiques/9025878)
- **S3** · CNIL / Conseil de l'IA et du numérique (CIANum), juillet 2026. Systèmes agentiques : traçabilité du processus décisionnel. https://www.cnil.fr/sites/default/files/2026-07/ia-cianum-cnil.pdf (https://www.cnil.fr/sites/default/files/2026-07/ia-cianum-cnil.pdf)
- **S4** · CNIL / Conseil de l'IA et du numérique (CIANum), juillet 2026. Systèmes agentiques : principe de minimisation. https://www.cnil.fr/sites/default/files/2026-07/ia-cianum-cnil.pdf (https://www.cnil.fr/sites/default/files/2026-07/ia-cianum-cnil.pdf)
- **S5** · CNIL / Conseil de l'IA et du numérique (CIANum), juillet 2026. Systèmes agentiques : risque de sécurité systémique. https://www.cnil.fr/sites/default/files/2026-07/ia-cianum-cnil.pdf (https://www.cnil.fr/sites/default/files/2026-07/ia-cianum-cnil.pdf)
- **S6** · NIST, Center for AI Standards and Innovation, décembre 2025 (publié en janvier 2025). Évaluation du détournement d'agents. https://www.nist.gov/node/1872381 (https://www.nist.gov/node/1872381)
- **S7** · NIST, CAISI Research Blog, mars 2026. Compétition de red teaming à grande échelle. https://www.nist.gov/node/1909361 (https://www.nist.gov/node/1909361)



PUBLIÉ SOUS LA RESPONSABILITÉ DE

Pierre-Louis Gaudreau, Président d'IA.consulting. Sources citées et datées. Une erreur ? Écrivez-nous.