



LIVRE BLANC

Guide du DSI sur le développement agentique : créer des applications d'entreprise à la vitesse de l'IA sans sacrifier la fiabilité

Guide à l'intention des COO,
CXO et VP Operations Leaders



La pression est réelle. Le problème aussi.

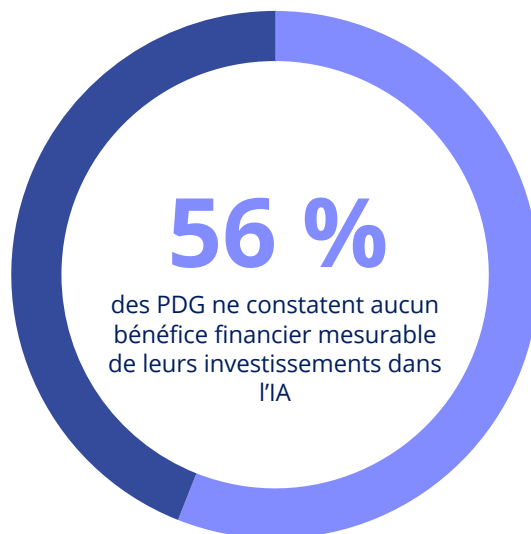
Les conseils d'administration ne se demandent plus s'il faut adopter l'IA. Ils veulent savoir quand, à quelle vitesse et avec quel retour sur investissement. Pour les COO et les responsables des opérations, la pression est forte. Les pilotes d'agents IA autonomes se multiplient. En effet, l'IA gère les interactions clients, achemine les dossiers, vérifie l'application des règles et pilote les workflows avec très peu d'intervention humaine.

Les premiers résultats semblaient prometteurs : délais de résolution plus courts, besoins en effectifs réduits et moins d'escalades manuelles. Mais le passage en production révèle une tout autre réalité :

les coûts ne suivent pas les prévisions, les résultats sont moins réguliers que ne le laissaient espérer les démos, et le contrôle opérationnel que les responsables pensaient conserver s'érode peu à peu.

56 % des PDG ne constatent aucun bénéfice financier mesurable de leurs investissements dans l'IA ([PwC 29th Global CEO Survey, janvier 2026](#)). Ce n'est pas un échec de l'IA. C'est un problème d'architecture. La question n'est donc plus de savoir si l'IA peut améliorer l'efficacité opérationnelle, mais plutôt de savoir si la façon dont les entreprises la déploient aujourd'hui leur permet réellement d'en tirer les bénéfices attendus.

Pour la plupart, la réponse est non.



Le coût caché du raisonnement

L'IA a profondément bouleversé l'économie de l'automatisation en entreprise. Les tokens, qui mesurent la quantité de « raisonnement » mobilisée par un système d'IA, représentent aujourd'hui un poste majeur des budgets de transformation des entreprises. Dans la plupart, le coût des tokens déjoue aujourd'hui toutes les prévisions.

Sur le papier, lorsque les dirigeants ont donné leur feu vert au déploiement d'agents IA, le modèle économique semblait pourtant simple : automatiser le raisonnement, réduire le travail manuel et préserver les marges. Mais un élément essentiel avait été sous-estimé : le passage à l'échelle rend la consommation de tokens extrêmement difficile à prévoir. Deux facteurs l'expliquent, et tous deux mettent à mal le ROI promis aux conseils d'administration.

Premier facteur : l'effet cumulatif de la consommation de tokens par les agents IA. Lorsqu'ils exécutent des tâches complexes, les agents IA ne se contentent pas d'un seul raisonnement avant de poursuivre. Ils recommencent à chaque étape, réexaminent le contexte, évaluent les options et génèrent de nouveaux résultats, encore et encore. La consommation s'accumule donc à deux niveaux : il faut des tokens pour créer et développer les systèmes, puis encore des tokens pour les faire fonctionner. À l'échelle des opérations d'une entreprise de taille moyenne, la facture peut alors grimper bien plus vite que ne le prévoyaient les modèles financiers.

C'est ce qu'on appelle la « taxe sur les tokens ».

Deuxième facteur : l'instabilité des prix. Les premiers fournisseurs d'IA proposaient des offres subventionnées et illimitées, ce qui permettait aux entreprises de prévoir les coûts. Mais cette époque est révolue. Pucés, énergie, capacité des data centers : les ressources sont sous tension. Parallèlement, les fournisseurs d'IA se tournent vers les marchés financiers. Par conséquent, les tarifs évoluent et les forfaits cèdent progressivement la place à la facturation à l'utilisation. Une même charge de travail n'a donc plus nécessairement le même prix. En juin 2026, Sam Altman reconnaissait publiquement que le coût des tokens était devenu « un énorme problème », y compris pour les fournisseurs eux-mêmes ([Axios, juin 2026](#)).

Pour les entreprises, les conséquences sont déjà visibles : les budgets de transformation s'épuisent sans que le ROI soit au rendez-vous. Gartner prévoit que plus de 40 % des projets d'IA agentique seront abandonnés d'ici fin 2027. Parmi les principales raisons : des coûts opérationnels en hausse et de plus en plus difficiles à prévoir ([Gartner, juin 2025](#)).

Pour les dirigeants, il ne s'agit donc plus de savoir s'il faut utiliser l'IA, mais comment la rentabiliser lorsque ses coûts ne cessent de fluctuer. Pour y parvenir, il faut changer de logique : ne plus payer pour des tokens, mais pour des résultats. Cela suppose de concevoir une architecture dans laquelle les dépenses en tokens d'IA sont réservées aux tâches qui exigent réellement du raisonnement. Tout le reste doit s'exécuter selon un modèle de coûts fixe, prévisible et sans effet cumulatif.

Quand l'IA sort du cadre

L'imprévisibilité des coûts n'est qu'une facette du problème. L'autre apparaît lorsque les agents IA se voient confier des décisions qu'ils ne devraient jamais avoir à prendre.

Les grands modèles de langage sont entraînés pour fournir des réponses utiles et fluides, pas pour respecter les politiques de conformité de votre entreprise. Ils ne connaissent pas par défaut vos SLA, vos obligations réglementaires, vos règles en matière d'escalade, ni les engagements précis pris auprès de vos clients: Sans garde-fous stricts dans les workflows en interaction avec les clients, un agent autonome peut produire une réponse qui lui semble parfaitement raisonnable, mais que votre équipe de conformité n'aurait jamais validée.

Ce n'est pas seulement une question de tolérance au risque. C'est un problème d'architecture.

Les décisions qui présentent les risques juridiques et réputationnels les plus élevés ont souvent une réponse correcte clairement définie : déterminer une éligibilité, résoudre un litige, appliquer une exception ou produire une communication réglementée. Ces décisions n'étant pas ambiguës, le recours à l'inférence probabiliste n'apporte aucune valeur ajoutée. Acheminer un dossier selon le statut d'un compte, vérifier qu'une demande respecte une liste de critères documentés ou appliquer une exception selon des conditions précises : toutes ces opérations sont déterministes. Les confier au raisonnement d'un agent IA n'apporte aucune valeur supplémentaire. Au contraire : cela introduit de la variabilité là où la constance est essentielle, et du risque là où la précision est indispensable.

À grande échelle, les hallucinations de l'IA ne sont pas de simples anomalies en production. Elles sont prévisibles : tout système qui génère des réponses par inférence probabiliste plutôt qu'à partir de règles déterministes peut en produire. Lorsque l'enjeu est faible, une hallucination reste un désagrément. Mais dans le cadre d'un litige client, un workflow de gestion des sinistres, une interaction relevant de services financiers ou une communication dans un secteur réglementé, elle devient un risque pour l'entreprise.

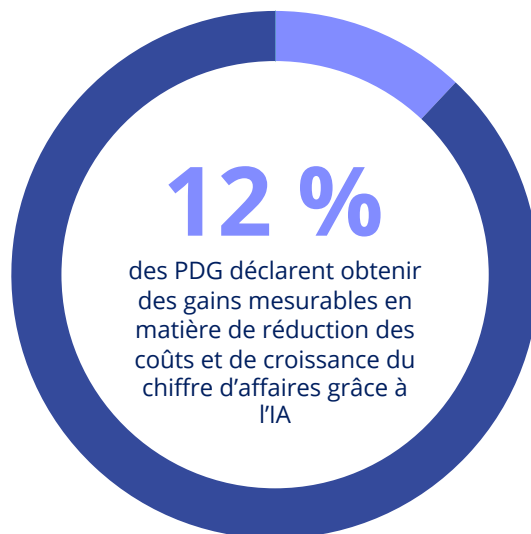
Gartner va plus loin. Parmi ses principales prévisions pour 2026, le cabinet estime que plus de 2 000 actions en justice liées à des décès imputables à l'IA seront engagées dans le monde. En cause : une gouvernance et des contrôles des risques insuffisants sur les résultats générés par l'IA ([Gartner, octobre 2025](#)). Ces risques ne sont ni théoriques ni lointains. Toute entreprise qui accorde aujourd'hui à des agents IA autonomes un réel pouvoir de décision dans des workflows réglementés ou en contact avec les clients y est déjà exposée.

La réponse doit donc venir de l'architecture. Les règles métier doivent déterminer ce que les agents IA peuvent et ne peuvent pas faire. L'IA ne doit jamais devenir le système de référence de votre logique de conformité, elle doit fonctionner dans le cadre imposé par cette logique. Pour les décisions sans ambiguïté, la logique métier doit déterminer l'action à suivre et la couche d'orchestration doit en garantir l'application. Il n'y a aucune raison de déléguer ces décisions à un modèle susceptible d'aboutir à une réponse différente d'un jour à l'autre. Cette séparation change profondément la donne. La conformité ne repose plus sur une supervision permanente : elle est intégrée à l'architecture même du système. L'IA ne peut plus sortir du cadre, car les règles ne peuvent pas être contournées par le raisonnement.

L'architecture qui répond aux deux enjeux

Le problème de la plupart des entreprises n'est pas d'utiliser trop d'IA, mais de ne pas l'utiliser au bon endroit dans l'architecture de leurs processus.

Pourtant, le potentiel économique de l'IA dans les opérations est bien réel : réduire le coût de service, augmenter la capacité de traitement et offrir une expérience client plus cohérente à grande échelle. Ces bénéfices sont à portée de main, mais pas en utilisant l'IA partout. Il faut savoir précisément où elle crée de la valeur et où, au contraire, elle ne fait qu'accroître les coûts et les risques sans retour à la hauteur.



Pour trouver le bon équilibre, l'architecture doit reposer sur deux couches complémentaires. La première est une couche d'orchestration déterministe : un moteur basé sur des règles qui pilote le déroulement des processus, applique la logique métier, suit leur état et achemine les tâches selon des critères clairement définis. Son coût marginal par transaction est proche de zéro. Elle ne raisonne pas, elle exécute. Son fonctionnement est entièrement prévisible et auditable, et elle n'hallucine pas.

La deuxième couche repose sur l'inférence par l'IA, qui intervient uniquement lorsque la logique déterministe ne suffit plus : analyser le ressenti client, interpréter des documents non structurés, générer des réponses en langage naturel ou synthétiser des dossiers complexes. Dans ces situations, le caractère probabiliste de l'IA devient un atout plutôt qu'un risque. Pourquoi ? Parce que ces tâches n'ont pas qu'une seule réponse correcte qu'un ensemble de règles pourrait déterminer.

Lorsque les rôles de ces deux couches sont clairement définis, la taxe sur les tokens chute de manière significative. Les tokens sont réservés aux tâches qui exigent réellement une inférence, et tout le reste s'exécute de manière déterministe. Cette séparation réduit aussi le risque de non-conformité, puisque la couche d'orchestration impose les règles métier, indépendamment de ce que produit l'IA. Celle-ci intervient donc dans un cadre strict, sans pouvoir contourner les règles par son propre raisonnement.

Seuls 12 % des PDG déclarent obtenir des gains mesurables sur les coûts et le chiffre d'affaires grâce à l'IA ([PwC, janvier 2026](#)). Ce qui distingue ces entreprises, ce n'est pas la quantité d'IA qu'elles déploient. C'est la rigueur avec laquelle elles l'intègrent à leur architecture.

Le contrôle : un avantage concurrentiel

Les responsables opérationnels qui rentabiliseront durablement leurs investissements dans l'IA ne seront pas ceux qui l'auront déployée le plus vite. Ni ceux qui auront donné le plus d'autonomie à leurs agents. Ce seront ceux qui auront conservé la maîtrise de l'architecture de leurs processus et intégré l'IA précisément là où elle crée de la valeur.

Cela implique de faire de l'orchestration un actif stratégique, et non une contrainte héritée du passé. Concrètement, il faut distinguer les décisions déterministes de celles qui nécessitent une inférence, puis concevoir une architecture qui garantit cette séparation. Pour que la taxe sur les tokens soit visible et ne pèse pas sur les budgets, les dépenses liées à l'IA doivent être mesurées au niveau de chaque transaction, et pas seulement à celui de la plateforme.

Le conseil d'administration a raison : l'IA peut réduire le coût de service et optimiser l'expérience client. Mais pour y parvenir, il ne suffit pas de laisser des agents autonomes déterminer eux-mêmes le fonctionnement des workflows. Il faut une architecture de processus gouvernée et orchestrée, dans laquelle l'IA intervient de manière ciblée.

Envie de prendre le contrôle sur vos dépenses ? Utilisez le [calculateur de TCO](#) pour comparer vos dépenses actuelles en tokens à celles d'une approche d'orchestration structurée, appliquée à vos workflows qui traitent les plus gros volumes.





Merci

À propos de Pega

Pega offre la plateforme basée sur l'IA leader pour accompagner la transformation des entreprises. Les organisations les plus influentes au monde font confiance à notre technologie pour repenser leur façon de travailler, avec l'automatisation des workflows, la personnalisation de l'expérience client et la modernisation des systèmes legacy. Depuis 1983, notre architecture évolutive et flexible favorise l'innovation continue et aide nos clients dans leur parcours vers l'entreprise autonome.

[PEGA.COM/FR](https://www.pega.com/fr)