



Le coût caché de l'IA ne vient pas du modèle, mais de la façon dont vous l'exploitez

Pour une IA prévisible

Juin 2026 | Don Schuerman, CTO et Head of Marketing



La prochaine phase de l'IA ne consiste plus à explorer le champ des possibles, mais à générer des résultats concrets et mesurables.

En deux ans, l'IA d'entreprise est passée du stade de l'expérimentation à la mise en pratique. Ce qui nécessitait autrefois des équipes spécialisées et de longs cycles de développement peut aujourd'hui être réalisé en quelques jours. Les équipes peuvent générer du code, créer des agents et automatiser des workflows à une vitesse spectaculaire.

Cette accélération a profondément changé les discussions au plus haut niveau.

Il y a un an, les dirigeants se demandaient : *que peut nous apporter l'IA ?*

Aujourd'hui, la question est plus directe : *combien nous coûte-t-elle et quelle valeur génère-t-elle réellement ?*

À mesure que les entreprises dépassent le stade des projets pilotes pour déployer l'IA à grande échelle, les exigences montent d'un cran. L'IA ne relève plus seulement de l'innovation. Elle devient partie intégrante des opérations, et doit, à ce titre, faire ses preuves en matière de fiabilité, de gouvernance et de coûts.

Dans tous les secteurs, un constat s'impose : si les promesses de l'IA sont claires, la réalité des coûts l'est beaucoup moins. Selon une récente étude Gartner¹, la plupart des initiatives d'IA peinent encore à générer des résultats tangibles. Pour de nombreuses entreprises, les premiers investissements produisent peu, voire aucun ROI mesurable.

Faut-il y voir un échec de l'IA ? Non. Nous entrons simplement dans une nouvelle phase : désormais, la réussite dépend moins de ce que les entreprises sont capables de créer que de leur capacité à exploiter l'IA efficacement à grande échelle.

¹ Étude Gartner sur la valeur de l'IA ([gartner.com/en/articles/ai-value](https://www.gartner.com/en/articles/ai-value))

Créer de l'IA n'a jamais été aussi simple, mais l'exploiter efficacement reste un défi

La première vague d'adoption de l'IA a été portée par sa démocratisation. Avec des modèles puissants désormais accessibles à tous, les entreprises ont multiplié les expérimentations et accéléré l'innovation. Mais le passage en production fait apparaître une tout autre réalité.

Les coûts, encore maîtrisables au stade pilote, augmentent rapidement. La qualité des résultats fluctue. Et surtout, il devient difficile d'établir un lien clair entre les investissements dans l'IA et la valeur réellement générée pour l'entreprise. Le problème est structurel : la plupart des systèmes d'IA actuels ont été conçus **à des fins d'expérimentation, et non pas pour fonctionner durablement.**



De nombreux systèmes s'appuient largement sur le *raisonnement à l'exécution* : le modèle repart de zéro à chaque interaction. Mais cette flexibilité a un prix. Chaque interaction mobilise de la puissance de calcul. Plus la tâche se complexifie, plus les ressources nécessaires augmentent. Et plus les usages se multiplient, plus la facture grimpe.

Réévaluer une tâche, interpréter son contexte, décider de l'étape suivante : chaque nouveau raisonnement de l'IA consomme des ressources supplémentaires. À petite échelle, l'impact reste maîtrisable. Mais lorsque les volumes augmentent, ces coûts s'additionnent rapidement. Le système ne se contente plus d'exécuter une tâche : il doit sans cesse refaire le même travail de réflexion. C'est ainsi que s'installe une véritable « taxe sur les tokens », où chaque étape de raisonnement supplémentaire vient alourdir la facture.

Et lorsque l'IA passe à l'échelle, la structure des coûts bascule. Dans de nombreux déploiements, le coût des inférences, c'est-à-dire de l'exécution continue des modèles en production, finit rapidement par dépasser celui de leur développement initial. L'IA n'est alors plus un investissement ponctuel, mais une dépense opérationnelle continue.

Le véritable coût de l'IA ne tient pas à ce que vous créez, mais à tout ce qu'il faut continuer à faire fonctionner. Les entreprises paient en permanence pour que l'IA détermine comment effectuer des tâches qui auraient pu être structurées en amont. C'est pourquoi des solutions très performantes dans l'environnement maîtrisé d'un pilote peuvent perdre en efficacité une fois déployées dans des conditions réelles. **Le problème n'est pas ce que l'IA peut faire, mais son efficacité.**

²MIT Technology Review, « AI's energy usage and climate footprint », 2025

Le véritable coût de l'IA dépend de la façon dont le travail est effectué

Pour maîtriser leurs dépenses liées à l'IA, les dirigeants se concentrent souvent sur les modèles. Quel fournisseur choisir ? Quelles fonctionnalités privilégier ? À quel prix ?

Mais la vraie question est ailleurs : **où l'IA intervient-elle dans vos opérations ? Et comment le travail est-il organisé ?**

Dans de nombreuses entreprises, l'IA s'est développée au fil des besoins. Chaque équipe a déployé ses propres modèles, agents et workflows, avec peu de coordination. Résultat : les outils se multiplient, les doublons aussi. Les mêmes décisions sont prises plusieurs fois, à différents endroits et parfois selon des logiques différentes.

Selon les prévisions, les charges de travail d'IA en entreprise pourraient consommer plus de 20 fois plus de tokens au cours des prochaines années³. À mesure que les agents se généralisent, les usages explosent, de même que les coûts. Car trop souvent, l'IA vient simplement se greffer sur les processus existants, sans en repenser le fonctionnement.

On gagne en efficacité à la marge, mais sans véritable transformation. Deux problèmes systémiques en découlent :

1. Les inefficacités perdurent. Si un processus est fragmenté ou inutilement complexe, l'IA ne fait pas disparaître cette complexité : elle l'accélère.
2. Trop de décisions sont prises au moment de l'exécution. Au lieu de définir en amont la manière dont le travail doit s'enchaîner, les entreprises laissent l'IA déterminer la prochaine étape en temps réel, interaction après interaction.

Si cette flexibilité peut sembler séduisante, elle n'est pas sans conséquences : plus de variabilité, des coûts plus élevés, et moins de maîtrise sur les résultats. Le problème est donc avant tout structurel :

- Un système bien conçu réduit les efforts nécessaires avant même l'exécution.
- Un système mal structuré génère des coûts à chaque exécution.

Déployer l'IA sans repenser l'organisation du travail entre les systèmes et les équipes entraîne une hausse des coûts à mesure que les usages augmentent, au lieu de permettre de les stabiliser avec le passage à l'échelle.

Voilà pourquoi, dans de nombreuses entreprises, les dépenses liées à l'IA augmentent plus vite que la valeur qu'elle génère.

³ Goldman Sachs, « AI agents forecast to boost tech cash flow as usage soars »

L'IA ne corrige pas les problèmes des processus défaillants : elle les amplifie

Une approche courante consiste à « greffer » l'IA sur les processus existants. L'entreprise cible un workflow, comme le service client, l'onboarding ou la gestion des sinistres, puis y intègre une couche d'IA pour l'améliorer. L'expérience devient plus conversationnelle, et le système plus adaptatif.

Mais en coulisses, rien ne change vraiment. Les systèmes restent multiples, les décisions fragmentées, et les étapes déconnectées. C'est alors à l'IA de faire le travail : elle interprète, oriente et s'adapte en temps réel pour assurer la continuité du processus.

À court terme, l'expérience peut s'améliorer. Mais les coûts, eux, ne diminuent pas. Ils peuvent même augmenter. Le système doit désormais effectuer le travail initial tout en mobilisant constamment les capacités d'interprétation de l'IA pour assurer la cohérence de l'ensemble.

Chaque fois que l'IA doit réévaluer la situation ou déterminer comment poursuivre le workflow, elle consomme des ressources supplémentaires. À l'échelle de milliers, voire de millions d'interactions, l'impact est considérable. L'IA devient alors une couche d'interprétation permanente qui vient compenser, sans la résoudre, la complexité des processus existants.

Ce modèle atteint rapidement ses limites à grande échelle. Pour maîtriser durablement les coûts, il faut d'abord repenser le processus lui-même.

La maîtrise des coûts commence dès la conception, pas par l'optimisation à l'exécution.

Les entreprises qui obtiennent des résultats tangibles grâce à l'IA ont changé d'approche. Elles privilégient un modèle orchestré, où l'IA est coordonnée, structurée et intégrée à la façon dont le travail est conçu et exécuté.

Plutôt que de laisser l'IA décider en temps réel de la marche à suivre, elles l'utilisent dès la conception des processus. C'est en amont qu'elle peut déployer tout son potentiel : analyser l'existant, identifier les inefficacités et repenser l'organisation du travail entre les systèmes et les équipes.

Une fois le processus clairement défini, l'exécution devient bien plus efficace. Le système n'a plus à déterminer chaque étape en temps réel : il suit un chemin prédéfini et fait appel à l'IA uniquement là où elle apporte une réelle valeur ajoutée, par exemple pour interpréter une intention ou traiter des données non structurées.

Ce changement de logique transforme à la fois les coûts et la fiabilité des résultats. En passant du *raisonnement à l'exécution* à *l'intelligence dès la conception*, les entreprises peuvent :

- réduire les calculs inutiles ;
- supprimer les prises de décision répétitives ;
- garantir une exécution reproductible et gouvernée.

L'IA n'a alors plus à piloter l'ensemble du processus : elle devient une composante d'un système d'exécution.

La prévisibilité ouvre la voie à l'IA à l'échelle de l'entreprise

Pour s'imposer comme un pilier de l'entreprise, l'IA doit répondre aux mêmes exigences de rigueur que tout autre système critique.

Et cette rigueur passe d'abord par la prévisibilité. Les responsables doivent pouvoir répondre à trois questions essentielles :

1. **Combien coûtera-t-elle ?**
2. **Quels résultats produira-t-elle ?**
3. **Comment se comportera-t-elle en conditions réelles ?**

Sans réponses claires, le passage à l'échelle devient risqué.

Or, selon une récente étude EY, plus de la moitié des projets d'IA sont déployés sans cadre de gouvernance. À mesure que l'usage se généralise, les risques s'amplifient : dépassements de coûts, non-conformité et résultats difficiles à prévoir.



C'est là qu'une approche plus structurée de l'IA fait toute la différence. Lorsque l'IA s'inscrit dans des workflows clairement définis, encadrés par l'orchestration et la gouvernance, les bénéfices sont concrets :

- Les résultats gagnent en cohérence.
- Les coûts sont plus prévisibles et plus directement liés aux décisions, aux transactions et aux résultats métier.
- Les opérations sont plus simples à piloter.

Mais la prévisibilité va bien au-delà de la consommation de tokens. Elle concerne l'ensemble du coût total de possession, des opérations à la gouvernance et à la conformité, en passant par les efforts nécessaires pour développer les systèmes au fil du temps. Le coût total de possession devient ainsi plus transparent. Il ne se limite plus aux tokens consommés, mais intègre le coût d'exploitation, de gouvernance, de maintenance et de passage à l'échelle de l'IA dans la durée.

Résultat : au lieu d'utiliser l'IA sans en définir clairement le cadre, les entreprises en font une composante maîtrisée et mesurable de leur fonctionnement. Elles passent ainsi du suivi de la consommation au suivi des résultats, une étape essentielle pour faire de l'IA un investissement réellement rentable.

⁴ EY, « Autonomous AI adoption surges as oversight falls behind », 2026

Pour réussir, les entreprises doivent repenser leur approche de l'IA.

De nombreuses entreprises pensent encore, à tort, devoir choisir entre accélérer avec l'IA et garder le contrôle. En réalité, l'enjeu n'est pas la vitesse, mais l'architecture.

Prendre une longueur d'avance ne consiste pas simplement à adopter l'IA plus vite que les autres.

Les entreprises qui en tireront pleinement parti seront celles qui repenseront leur fonctionnement. Concrètement, cela implique de revoir les priorités :

Passer des outils aux systèmes

La vraie question n'est pas de savoir ce que l'IA peut accomplir seule, mais comment elle s'intègre aux workflows de l'entreprise.

Repenser avant d'automatiser

L'IA ne suffit pas à corriger les failles d'un processus. Pour produire un réel impact, il faut d'abord repenser le processus lui-même.

Mesurer les résultats, pas l'activité

La réussite doit se mesurer aux résultats obtenus pour l'entreprise : tâches accomplies, expériences améliorées, obstacles réduits, et non au simple volume d'utilisation.

Faire de l'IA une infrastructure opérationnelle

À mesure que l'IA s'intègre aux processus clés de l'entreprise, elle doit répondre aux mêmes exigences de fiabilité et de gouvernance que les autres systèmes.

L'objectif n'est plus d'adopter de nouvelles technologies, mais de mieux maîtriser la façon dont elles sont utilisées.

De l'expérimentation à l'exécution

L'IA d'entreprise est à un tournant. La première phase a révélé tout son potentiel. La prochaine devra démontrer ce qui peut réellement fonctionner dans la durée et à grande échelle.

À mesure que les entreprises multiplient les usages de l'IA, le défi change de nature : il ne s'agit plus simplement de produire des insights ou d'automatiser des tâches isolées, mais d'exploiter l'IA de façon fiable et efficace, au service de résultats métier concrets.

C'est ce qui permet de distinguer les succès ponctuels d'une véritable création de valeur à l'échelle de l'entreprise.

Les entreprises qui réussiront cette transition pourront gagner en rapidité et en efficacité opérationnelle, et s'adapteront mieux au changement..

Mais déployer toujours plus d'IA ne suffira pas.

Pour en tirer pleinement parti, elles devront intégrer durablement l'IA au cœur de leurs opérations.

En résumé

L'IA transforme déjà ce que les entreprises sont capables d'accomplir.

Le véritable défi commence maintenant : faire ses preuves en conditions réelles, à grande échelle, dans des environnements complexes, tout en maîtrisant les coûts et en gardant le contrôle.

Car le principal coût caché de l'IA ne vient pas du modèle lui-même ; il réside dans la manière dont le travail est organisé, exécuté et répété dans le temps.

Les entreprises qui sauront relever ce défi ne réduiront pas seulement les coûts ; elles disposeront d'une infrastructure capable de soutenir durablement leurs opérations.

Et dans la prochaine ère de l'IA d'entreprise, l'avantage ira à celles qui auront su en faire une technologie maîtrisée, mesurable et pleinement intégrée à leurs opérations.



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)