Saisir l'avantage de l'IA agentive

Un guide pratique pour les PDG afin de résoudre le paradoxe de l'IA générationnelle et de débloquer un impact à grande échelle grâce aux agents d'IA.

En un coup d'œil

- Près de huit entreprises sur dix déclarent utiliser l'IA de nouvelle génération, mais tout autant ne constatent aucun impact significatif sur leurs résultats financiers. **Considérez cela comme le « paradoxe de l'IA générationnelle ».**
- Au cœur de ce paradoxe se trouve un déséquilibre entre les copilotes et les chatbots « horizontaux » (à l'échelle de l'entreprise) — qui se sont développés rapidement mais n'offrent que des gains diffus et difficiles à mesurer — et les cas d'utilisation « verticaux » (spécifiques à une fonction) plus transformateurs — dont environ 90 % restent bloqués en mode pilote**.**
- Les agents d'IA offrent une solution au paradoxe de l'IA généraliste. En effet, ils ont le potentiel d'automatiser des processus métier complexes — en combinant autonomie, planification, mémoire et intégration — transformant ainsi l'IA généraliste d'un outil réactif en un collaborateur virtuel proactif et orienté vers un objectif.
- Les agents d'IA offrent une solution au paradoxe de l'IA généraliste. En effet, ils ont le potentiel d'automatiser des processus métier complexes — en combinant autonomie, planification, mémoire et intégration — transformant ainsi l'IA généraliste d'un outil réactif en un collaborateur virtuel proactif et orienté vers un objectif.
- Les agents d'IA offrent une solution au paradoxe de l'IA généraliste. En effet, ils ont le potentiel d'automatiser des processus métier complexes — en combinant autonomie, planification, mémoire et intégration — transformant ainsi l'IA généraliste d'un outil réactif en un collaborateur virtuel proactif et orienté vers un objectif.
- .Ce changement permet bien plus qu'une simple amélioration de l'efficacité. Les agents décuplent l'agilité opérationnelle et créent de nouvelles sources de revenus.
- Mais pour exploiter pleinement le potentiel de l'IA agentielle, il ne suffit pas d'intégrer des agents aux flux de travail existants. Il faut repenser ces flux de travail de A à Z, en plaçant les agents au cœur du système.

Avant-propos

- Un nouveau paradigme d'architecture d'IA – le maillage d'IA agentique – est nécessaire pour encadrer l'évolution rapide du paysage de l'IA organisationnelle et permettre aux équipes de combiner agents sur mesure et agents prêts à l'emploi, tout en maîtrisant la dette technique croissante et les nouveaux types de risques. Mais le plus grand défi ne sera pas technique. Il sera humain : gagner la confiance, favoriser l'adoption et

mettre en place une gouvernance adéquate pour gérer l'autonomie des agents et prévenir leur prolifération incontrôlée.

- Mais pour exploiter pleinement le potentiel de l'IA agentielle, il ne suffit pas d'intégrer des agents aux flux de travail existants. Il faut repenser ces flux de travail de A à Z, en plaçant les agents au cœur du système.
- Pour amplifier leur impact à l'ère de l'intelligence artificielle, les organisations doivent repenser leurs approches de transformation par l'IA : passer d'initiatives dispersées à des programmes stratégiques ; des cas d'utilisation aux processus métier ; des équipes d'IA cloisonnées aux équipes de transformation transversales ; et de l'expérimentation à une mise en œuvre industrialisée et évolutive.
- Les organisations devront également mettre en place les fondements nécessaires pour opérer efficacement à l'ère des agents. Elles devront développer les compétences de leurs employés, adapter leur infrastructure technologique, accélérer la valorisation des données et déployer des mécanismes de gouvernance spécifiques aux agents. Le moment est venu de clore le chapitre de l'expérimentation en IA générale — un tournant que seul le PDG peut opérer.

Le paradoxe de l'IA générationnelle : déploiement généralisé, impact minimal

Points clés

- Près de huit entreprises sur dix ont déployé l'IA de nouvelle génération sous une forme ou une autre, mais à peu près le même pourcentage déclare qu'elle n'a eu aucun impact significatif sur ses bénéfices. Nous appelons cela le « **paradoxe de l'IA générale** ».
- Le principal problème réside dans le déséquilibre entre les cas d'usage « horizontaux » et « verticaux ». Les premiers, comme les copilotes employés et les chatbots, sont largement déployés mais leurs bénéfices restent diffus, tandis que les cas d'usage verticaux, ou spécifiques à une fonction, à plus fort impact, dépassent rarement le stade pilote en raison d'obstacles techniques, organisationnels, liés aux données et culturels.
- À moins que les entreprises ne s'attaquent à ces obstacles, la promesse transformationnelle de l'IA de nouvelle génération restera largement inexploitée.

L'intelligence artificielle de nouvelle génération est partout, sauf dans les comptes de résultat des entreprises

Avant même l'avènement de l'IA de nouvelle génération, l'intelligence artificielle s'était déjà taillé une place essentielle dans les entreprises, en alimentant des capacités avancées de prédiction, de classification et d'optimisation. Son potentiel de valeur était déjà estimé à un montant immense, entre 11 et 18 billions de dollars à l'échelle mondiale.

Principalement dans les domaines du marketing (permettant des fonctionnalités telles que le ciblage personnalisé des e-mails et la segmentation client), des ventes (qualification des prospects) et de la chaîne d'approvisionnement (optimisation des stocks et prévision de la demande). Pourtant, l'IA restait largement l'apanage des experts. De ce fait, son adoption par l'ensemble des employés était généralement lente. De 2018 à 2022, par exemple, l'adoption de l'IA est restée relativement stable, environ 50 % des entreprises déployant cette technologie dans une seule fonction métier, selon une étude de McKinsey

L'IA de nouvelle génération a étendu la portée de l'IA traditionnelle dans trois domaines révolutionnaires : la synthèse d'informations, la génération de contenu et la communication en langage humain. McKinsey estime que cette technologie pourrait générer entre 2 600 et 4 400 milliards de dollars de valeur ajoutée, en plus de celle déjà présente dans l'IA analytique traditionnelle.3

Deux ans et demi après le lancement de ChatGPT, l'IA nouvelle génération a profondément transformé la manière dont les entreprises interagissent avec l'IA. Son potentiel de transformation réside non seulement dans les nouvelles fonctionnalités qu'elle introduit, mais aussi dans sa capacité à démocratiser l'accès aux technologies d'IA avancées au sein des organisations. Cette démocratisation a entraîné une forte augmentation de la sensibilisation à l'IA et des expérimentations associées : selon la dernière enquête mondiale de McKinsey sur l'IA ,4Plus de 78 % des entreprises utilisent désormais l'IA de nouvelle génération dans au moins une fonction commerciale (contre 55 % un an auparavant).

Toutefois, cet enthousiasme ne s'est pas encore traduit par des résultats économiques concrets. Plus de 80 % des entreprises déclarent toujours que leurs initiatives en matière d'IA de nouvelle génération n'ont aucun impact significatif sur leurs bénéfices.5De plus, seulement 1 % des entreprises que nous avons interrogées considèrent leurs stratégies en matière d'IA de génération comme matures .6On pourrait parler du « paradoxe de l'IA générationnelle » : malgré toute l'énergie, les investissements et le potentiel entourant cette technologie, son impact à grande échelle ne s'est pas encore concrétisé pour la plupart des organisations.

Au cœur du paradoxe de l'IA générationnelle se trouve un déséquilibre entre les cas d'utilisation horizontaux et verticaux.

De nombreuses organisations ont déployé des cas d'utilisation horizontaux, tels que des copilotes et des chatbots à l'échelle de l'entreprise ; près de 70 % des entreprises du classement Fortune 500, par exemple, utilisent Microsoft 365 Copilot.7Ces outils sont généralement perçus comme des leviers d'amélioration de la productivité individuelle, permettant aux employés de gagner du temps sur les tâches routinières et d'accéder à l'information et de la synthétiser plus efficacement. Toutefois, ces améliorations, bien que réelles, ont tendance à être peu visibles parmi les employés. Par conséquent, elles ne se traduisent pas facilement par des résultats concrets, que ce soit au niveau du chiffre d'affaires ou des bénéfices.

À l'inverse, les cas d'usage verticaux — ceux intégrés à des fonctions et processus métier spécifiques — ont connu un déploiement limité dans la plupart des entreprises, malgré leur potentiel plus élevé d'impact économique direct (voir graphique 2). Selon une étude de McKinsey, moins de 10 % des cas d'usage déployés dépassent le stade du projet pilote .8Même une fois pleinement déployées, ces solutions n'ont généralement pris en charge que des étapes isolées d'un processus métier et ont fonctionné de manière réactive, sur intervention humaine, plutôt que de façon proactive ou autonome. De ce fait, leur impact sur la performance de l'entreprise est resté limité.

Comment expliquer ce déséquilibre ? D'une part, les solutions de copilotage déployées horizontalement, telles que Microsoft Copilot ou Google AI Workspace, sont accessibles et prêtes à l'emploi, et relativement faciles à mettre en œuvre. (Dans de nombreux cas, activer Microsoft Copilot se résume à ajouter une extension à un contrat Office 365 existant, sans nécessiter de refonte des flux de travail ni d'efforts majeurs de gestion du changement.) Le déploiement rapide des chatbots d'entreprise a également été motivé par des préoccupations liées à la réduction des risques. Face à l'expérimentation par les employés de modèles de langage externes de grande taille (LLM) comme ChatGPT, de nombreuses organisations ont mis en place des alternatives internes et sécurisées afin de limiter les fuites de données et de garantir la conformité aux politiques de sécurité de l'entreprise.

Repenser l'entreprise grâce à la technologie et à l'IA : une transformation radicale.

Rejoignez Robert Levin et Kate Smaje , associés principaux de McKinsey et co-auteurs de « Rewired : The McKinsey Playbook on How Leading Companies Win with Technology and AI », pour une conversation sur ce que font les entreprises leaders pour transformer leurs activités grâce à l'IA.*

Le déploiement limité et la portée restreinte des cas d'utilisation verticaux peuvent être attribués à six facteurs principaux :

- Initiatives fragmentées. Dans de nombreuses entreprises, les cas d'usage verticaux ont été identifiés par une approche ascendante et très granulaire au sein de chaque fonction. De fait, moins de 30 % des entreprises indiquent que leur PDG pilote directement leur stratégie en matière d'IA .9Cela a entraîné une prolifération de micro-initiatives déconnectées et une dispersion des investissements en IA, avec une coordination limitée au niveau de l'entreprise.*
- **Absence de solutions éprouvées et prêtes à l'emploi**. Contrairement aux applications horizontales standard, telles que les copilotes, les cas d'usage verticaux nécessitent souvent un développement sur mesure. De ce fait, les équipes sont fréquemment contraintes de tout construire de A à Z, en utilisant des technologies émergentes et en constante évolution avec lesquelles elles ont une expérience limitée. Si de nombreuses entreprises ont investi dans des data scientists pour développer des modèles d'IA, elles manquent souvent d'ingénieurs MLOps, pourtant essentiels pour industrialiser, déployer et maintenir ces modèles en production.
- **Limitations technologiques des LLM**. Malgré leurs capacités impressionnantes, la première génération de LLM présentait des limitations qui ont considérablement freiné leur déploiement à grande échelle. Premièrement, les LLM peuvent produire des résultats inexacts, ce qui les rend peu fiables dans les environnements où la précision et la reproductibilité sont essentielles. De plus, malgré leur puissance, les LLM sont fondamentalement passifs ; ils n’agissent que sur commande et ne peuvent pas gérer les flux de travail de manière autonome ni prendre de décisions sans intervention humaine. Les LLM ont également éprouvé des difficultés à gérer des

flux de travail complexes comportant plusieurs étapes, points de décision ou logiques de branchement. Enfin, de nombreux LLM actuels disposent d'une mémoire persistante limitée, ce qui rend difficile le suivi du contexte dans le temps ou un fonctionnement cohérent lors d'interactions prolongées.

- **Équipes d'IA cloisonnées**. Les centres d'excellence en IA ont joué un rôle crucial dans l'accélération de la sensibilisation et de l'expérimentation au sein de nombreuses organisations. Cependant, dans bien des cas, ces équipes ont fonctionné en silos, développant des modèles d'IA indépendamment des fonctions informatiques, de données ou métiers essentielles. Cette autonomie, bien qu'utile pour le prototypage rapide, a souvent rendu les solutions difficiles à déployer à grande échelle en raison d'une mauvaise intégration aux systèmes d'entreprise, de flux de données fragmentés ou d'un manque d'alignement opérationnel.
- **Lacunes en matière d'accessibilité et de qualité des données**. Ces lacunes concernent aussi bien les données structurées que non structurées, ces dernières restant largement non réglementées dans la plupart des organisations.
- **Craintes culturelles et inertie organisationnelle**. Dans de nombreuses organisations, le déploiement de l'IA s'est heurté à une résistance implicite de la part des équipes opérationnelles et du management intermédiaire, en raison de la crainte de perturbations, de l'incertitude quant à l'impact sur l'emploi et du manque de familiarité avec cette technologie.

Malgré son impact limité sur les résultats financiers jusqu'à présent, la première vague d'IA de génération s'est avérée loin d'être vaine. Elle a enrichi les compétences des employés, permis une expérimentation à grande échelle, accéléré la familiarisation avec l'IA dans tous les services et aidé les organisations à développer des compétences essentielles en ingénierie rapide, en évaluation des modèles et en gouvernance. Tout cela a jeté les bases d'une seconde phase plus intégrée et transformatrice : l'avènement des agents d'IA .

Chapitre 2

Du paradoxe à la récompense : comment les agents peuvent déployer l'IA à grande échelle

Points clés

- En automatisant les processus métier complexes, les agents permettent d'exploiter pleinement le potentiel des cas d'usage verticaux. Les entreprises visionnaires tirent déjà parti de la puissance des agents pour transformer leurs processus clés
- Pour exploiter pleinement le potentiel des agents, les entreprises doivent réinventer leur façon de travailler : modifier les flux de tâches, redéfinir les rôles humains et construire dès le départ des processus centrés sur les agents.
- Pour y parvenir, il faudra un nouveau paradigme d'architecture d'IA : le maillage d'IA agentif, capable d'intégrer des agents développés sur mesure et des agents prêts à l'emploi. Mais le plus grand défi ne sera pas technique. Il sera humain : gagner la confiance pour favoriser l'adoption et établir les protocoles de gouvernance appropriés.

La percée : L'automatisation des flux de travail complexes permet d'exploiter pleinement le potentiel des cas d'utilisation verticaux.

Les langages de modélisation linguistique (LLM) ont révolutionné la manière dont les organisations interagissent avec les données, permettant la synthèse d'informations, la génération de contenu et l'interaction en langage naturel. Cependant, malgré leur puissance, les LLM sont restés fondamentalement réactifs et isolés des systèmes d'entreprise, incapables de conserver en mémoire les interactions passées ou le contexte entre les sessions ou les requêtes. Leur rôle s'est longtemps limité à l'amélioration de la productivité individuelle par le biais de tâches isolées. Les agents d'IA marquent une évolution majeure de l'IA d'entreprise, faisant passer l'IA générative de la génération réactive de contenu à une exécution autonome et orientée vers un objectif. Ces agents peuvent comprendre les objectifs, les décomposer en sous-tâches, interagir avec les humains et les systèmes, exécuter des actions et s'adapter en temps réel, le tout avec une intervention humaine minimale. Ils y parviennent en combinant les LLM avec des composants technologiques supplémentaires offrant des capacités de mémoire, de planification, d'orchestration et d'intégration.

Grâce à ces nouvelles fonctionnalités, les agents d'IA étendent le potentiel des solutions horizontales, transformant les copilotes généralistes d'outils passifs en collaborateurs proactifs. Ces derniers ne se contentent plus de répondre aux sollicitations, mais surveillent également les tableaux de bord, déclenchent des flux de travail, assurent le suivi des actions en cours et fournissent des informations pertinentes en temps réel. Toutefois, la

véritable avancée réside dans le domaine vertical, où l'IA agentique permet l'automatisation de flux de travail métier complexes impliquant de multiples étapes, acteurs et systèmes – des processus qui dépassaient auparavant les capacités des outils d'IA de première génération.

Les agents apportent bien plus que de l'efficacité : ils décuplent l'agilité opérationnelle et ouvrent de nouvelles perspectives de revenus.

Côté opérations, les agents prennent en charge les tâches routinières et gourmandes en données, permettant ainsi aux humains de se concentrer sur des missions à plus forte valeur ajoutée. Mais leur rôle ne s'arrête pas là : ils transforment les processus de cinq manières :

- **Les agents accélèrent l'exécution en éliminant les délais entre les tâches et en permettant le traitement parallèle.** Contrairement aux flux de travail traditionnels qui reposent sur des transferts séquentiels, les agents peuvent coordonner et exécuter plusieurs étapes simultanément, réduisant ainsi le temps de cycle et améliorant la réactivité.
- **Les agents apportent l'adaptabilité**. En ingérant des données en continu, ils peuvent ajuster les flux de processus en temps réel, en réorganisant les séquences de tâches, en réattribuant les priorités ou en signalant les anomalies avant qu'elles ne provoquent des défaillances. Les flux de travail sont ainsi non seulement plus rapides, mais aussi plus intelligents.
- **Les agents permettent une personnalisation**. En adaptant les interactions et les décisions aux profils ou comportements individuels des clients, les agents peuvent ajuster le processus de manière dynamique afin de maximiser la satisfaction et les résultats.
- **Les agents apportent de la flexibilité aux opérations**. Étant numériques, leur capacité d'exécution peut augmenter ou diminuer en temps réel en fonction de la charge de travail, de la saisonnalité de l'activité ou des pics d'activité imprévus – une flexibilité difficile à obtenir avec des modèles de ressources humaines fixes.
- **Les agents contribuent également à renforcer la résilience des opérations**. En surveillant les perturbations, en réacheminant les opérations et en intervenant uniquement en cas de besoin, ils

assurent la continuité des processus, qu'il s'agisse de chaînes d'approvisionnement confrontées à des retards portuaires ou de flux de travail de service s'adaptant aux pannes de système.

Dans un environnement de chaîne d'approvisionnement complexe, un agent d'IA pourrait, par exemple, jouer le rôle d'une couche d'orchestration autonome pour les opérations d'approvisionnement, d'entreposage et de distribution. Connecté aux systèmes internes (tels que le système de planification de la chaîne d'approvisionnement ou le système de gestion d'entrepôt) et aux sources de données externes (comme les prévisions météorologiques, les flux de données des fournisseurs et les signaux de la demande), cet agent pourrait prévoir la demande en continu. Il pourrait ensuite identifier les risques, tels que les retards ou les perturbations, et replanifier dynamiquement les flux de transport et de stocks. En sélectionnant le mode de transport optimal en fonction du coût, du délai et de l'impact environnemental, l'agent pourrait réallouer les stocks entre les entrepôts, négocier directement avec les systèmes externes et remonter les décisions nécessitant une expertise stratégique. Résultat : une amélioration des niveaux de service, une réduction des coûts logistiques et une diminution des émissions.

Les agents peuvent également contribuer à stimuler la croissance du chiffre d'affaires en amplifiant les sources de revenus existantes et en en débloquant de nouvelles :

- **Augmenter les revenus existants**. Dans le e-commerce, des agents intégrés aux boutiques en ligne ou aux applications pourraient analyser proactivement le comportement des utilisateurs, le contenu de leur panier et le contexte (par exemple, la saisonnalité ou l'historique d'achats) afin de proposer des offres de vente incitative et de vente croisée en temps réel. Dans le secteur financier, ces agents pourraient aider les clients à trouver des produits financiers adaptés à leurs besoins, tels que des prêts, des assurances ou des portefeuilles d'investissement, en leur fournissant des conseils personnalisés en fonction de leur profil financier, de leur situation personnelle et de leurs habitudes de consommation.
- **Création de nouvelles sources de revenus**. Pour les entreprises industrielles, des agents intégrés aux produits ou équipements connectés pourraient surveiller l'utilisation, détecter les seuils de performance et débloquer automatiquement des fonctionnalités ou déclencher des interventions de maintenance, permettant ainsi des modèles de rémunération à l'usage, par abonnement ou basés sur la performance. De même, les entreprises de services pourraient

intégrer leur expertise interne (raisonnement juridique, interprétation fiscale et bonnes pratiques d'approvisionnement) dans des agents d'IA proposés sous forme de logiciels en tant que service (SaaS) ou d'API à leurs clients, partenaires ou PME ne disposant pas de cette expertise en interne.

En résumé, l'IA agentielle ne se contente pas d'automatiser. Elle redéfinit la manière dont les organisations fonctionnent, s'adaptent et créent de la valeur.

Ce n'est plus de la science-fiction : les entreprises visionnaires exploitent le pouvoir des agents

Les études de cas suivantes démontrent comment **QuantumBlack** aide les organisations à constituer des effectifs d'agents, avec des résultats qui vont bien au-delà des gains d'efficacité.

Étude de cas 1 : Comment une banque a utilisé des « usines numériques » hybrides pour moderniser ses applications existantes

Le problème : une grande banque devait moderniser son système central informatique, composé de 400 logiciels – un projet colossal dont le budget dépassait les 600 millions de dollars. De grandes équipes de développeurs ont abordé le projet en effectuant des tâches manuelles et répétitives, ce qui a engendré des difficultés de coordination entre les différents services. Elles s’appuyaient également sur une documentation et un code souvent lents et sujets aux erreurs. Si les outils d’IA de première génération ont permis d’accélérer certaines tâches, la progression est restée lente et laborieuse.

L'approche par agents : les employés humains ont été promus à des rôles de supervision, encadrant des équipes d'agents d'IA. Chaque équipe contribue à un objectif commun selon une séquence définie (voir l'illustration 3). Ces équipes documentent a posteriori l'application existante, écrivent du nouveau code, examinent le code des autres agents et intègrent ce code dans des fonctionnalités qui sont ensuite testées par d'autres agents avant la livraison du produit final. Libérés des tâches manuelles répétitives, les superviseurs humains guident chaque étape du processus, améliorant ainsi la qualité des livrables et réduisant le nombre de sprints nécessaires à l'implémentation de nouvelles fonctionnalités.

Impact : Réduction de plus de 50 % du temps et des efforts consacrés par les équipes pionnières.

Étude de cas 2 : Comment un cabinet d'études a amélioré la qualité de ses données pour obtenir des informations plus approfondies sur le marché

Le problème : une société d'études de marché et de veille stratégique consacrait des ressources considérables à garantir la qualité des données, s'appuyant sur une équipe de plus de 500 personnes chargées de collecter, structurer et codifier les données, puis de générer des analyses personnalisées pour ses clients. Ce processus, réalisé manuellement, était sujet aux erreurs, dont 80 % étaient détectées par les clients eux-mêmes.

L'approche multi-agents : une solution autonome identifie les anomalies de données et explique les variations des ventes ou des parts de marché. Elle analyse les signaux internes, tels que les modifications de la nomenclature des produits, et les événements externes identifiés par des recherches web, comme les rappels de produits ou les intempéries. Les facteurs les plus influents sont synthétisés, hiérarchisés et mis à la disposition des décideurs. Grâce à la recherche avancée et au raisonnement contextuel, les agents font souvent émerger des informations qu'il serait difficile pour des analystes humains de découvrir manuellement. Bien que le système ne soit pas encore en production, il est pleinement fonctionnel et a démontré un fort potentiel pour libérer les analystes de leurs tâches stratégiques.

Impact : Gain de productivité potentiel de plus de 60 % et économies attendues de plus de 3 millions de dollars par an.

Étude de cas 3 : Comment une banque a repensé sa méthode de création de notes d'évaluation des risques de crédit

Le problème : les chargés de clientèle d'une banque de détail passaient des semaines à rédiger et à peaufiner des notes d'évaluation du risque de crédit afin de faciliter leurs décisions et de se conformer aux exigences réglementaires (voir l'annexe 4). Ce processus les obligeait à examiner et à extraire manuellement des informations provenant d'au moins dix sources de données différentes, et à élaborer un raisonnement complexe et nuancé sur

des sections interdépendantes, par exemple l'évolution conjointe des prêts, des revenus et de la trésorerie.

L'approche par agents : En étroite collaboration avec les experts en risque de crédit et les chargés de clientèle de la banque, une preuve de concept a été développée afin de transformer le flux de travail des notes de crédit grâce à des agents d'IA. Ces agents assistent les chargés de clientèle en extrayant des données, en rédigeant des sections de notes, en générant des scores de confiance pour prioriser les analyses et en suggérant des questions de suivi pertinentes. Dans ce modèle, le rôle de l'analyste évolue de la rédaction manuelle vers la supervision stratégique et la gestion des exceptions.

Impact : Une augmentation potentielle de la productivité de 20 à 60 %, dont une amélioration de 30 % du délai de traitement des crédits

Pour maximiser la valeur des agents d'IA, il faut réinventer les processus.

Pour exploiter pleinement le potentiel de l'IA dans les secteurs verticaux, il ne suffit pas d'intégrer des agents aux flux de travail existants. Il est nécessaire de repenser entièrement la conception, en passant de l'automatisation des tâches au sein d'un processus existant à la réinvention du processus dans son intégralité, avec la collaboration d'humains et d'agents. En effet, lorsque des agents sont intégrés à un processus existant sans refonte, ils servent généralement d'assistants plus rapides : ils génèrent du contenu, extraient des données ou exécutent des étapes prédéfinies. Mais le processus lui-même reste séquentiel, rigide et soumis aux contraintes humaines.

Repenser un processus autour d'agents ne se limite pas à automatiser les flux de travail existants ; cela implique de repenser entièrement l'architecture du flux de tâches. Il s'agit notamment de réorganiser les étapes, de redistribuer les responsabilités entre humains et agents, et de concevoir le processus pour exploiter pleinement les atouts de l'IA agentielle : exécution parallèle réduisant considérablement les délais, adaptabilité en temps réel aux conditions changeantes, personnalisation poussée à grande échelle et capacité flexible s'ajustant instantanément à la demande.

Prenons l'exemple d'un centre d'appels clients hypothétique. Avant l'introduction d'agents IA, ce centre utilisait des outils d'IA générale pour assister le personnel d'assistance humain : recherche d'articles dans les bases de connaissances, synthèse de l'historique des tickets et aide à la rédaction des réponses. Si cette assistance a permis d'accélérer le processus et de réduire la charge cognitive, le processus lui-même restait

entièrement manuel et réactif, les agents humains gérant toujours chaque étape du diagnostic, de la coordination et de la résolution. Le potentiel d'amélioration de la productivité était modeste, avec généralement une augmentation du temps de résolution et de la productivité de 5 à 10 %.

Imaginez maintenant qu'un centre d'appels intègre des agents IA tout en conservant en grande partie son flux de travail actuel : des agents sont ajoutés pour intervenir à des étapes spécifiques sans que l'acheminement, le suivi ou la résolution des demandes ne soient modifiés de bout en bout. Ces agents peuvent classer les tickets, suggérer les causes probables, proposer des solutions et même résoudre de manière autonome les problèmes fréquents et simples (comme la réinitialisation des mots de passe). Si l'impact peut être accru – avec un gain de temps estimé entre 20 et 40 % et une réduction du volume de demandes en attente de 30 à 50 % –, les difficultés de coordination et la faible adaptabilité empêchent des gains véritablement significatifs.

Mais le véritable changement s'opère au troisième niveau, lorsque le processus du centre d'appels est repensé autour de l'autonomie des agents. Dans ce modèle, les agents IA ne se contentent pas de répondre : ils détectent proactivement les problèmes clients courants (tels que les retards de livraison, les échecs de paiement ou les interruptions de service) en analysant les tendances sur tous les canaux, anticipent les besoins probables, initient automatiquement les solutions (comme les remboursements, les commandes de nouveaux articles ou la mise à jour des informations de compte) et communiquent directement avec les clients par chat ou e-mail. Les agents humains sont repositionnés comme gestionnaires d'escalade et responsables de la qualité du service, et interviennent uniquement lorsque les agents détectent une incertitude ou une exception aux schémas habituels. L'impact à ce niveau est transformateur. Cela pourrait permettre une amélioration radicale de la productivité du service client. Jusqu'à 80 % des incidents courants pourraient être résolus de manière autonome, avec une réduction du temps de résolution de 60 à 90 % (voir l'illustration

.

Saisir l'avantage de l'IA agentive

13 juin 2025| Rapport

Share

Print

Download

Sauvegarder

Un guide pratique pour les PDG afin de résoudre le paradoxe de l'IA générationnelle et de débloquer un impact à grande échelle grâce aux agents d'IA.

TÉLÉCHARGEMENTS

Rapport complet (28 pages)

En un coup d'œil

Chapitre 1

Chapitre 2

Chapitre 3

Conclusion

En un coup d'œil

Près de huit entreprises sur dix déclarent utiliser l'IA de nouvelle génération, mais tout autant ne constatent aucun impact significatif sur leurs résultats financiers.1Considérez cela comme le « paradoxe de l'IA générationnelle ».

Signature

À propos des auteurs

Au cœur de ce paradoxe se trouve un déséquilibre entre les copilotes et les chatbots « horizontaux » (à l'échelle de l'entreprise) — qui se sont développés rapidement mais n'offrent que des gains diffus et difficiles à mesurer — et les cas d'utilisation « verticaux » (spécifiques à une fonction) plus transformateurs — dont environ 90 % restent bloqués en mode pilote.

Les agents d'IA offrent une solution au paradoxe de l'IA généraliste. En effet, ils ont le potentiel d'automatiser des processus métier complexes — en combinant autonomie, planification, mémoire et intégration — transformant

ainsi l'IA généraliste d'un outil réactif en un collaborateur virtuel proactif et orienté vers un objectif.

Ce changement permet bien plus qu'une simple amélioration de l'efficacité. Les agents décuplent l'agilité opérationnelle et créent de nouvelles sources de revenus.

Mais pour exploiter pleinement le potentiel de l'IA agentielle, il ne suffit pas d'intégrer des agents aux flux de travail existants. Il faut repenser ces flux de travail de A à Z, en plaçant les agents au cœur du système.

Partager

barre latérale

Avant-propos

Un nouveau paradigme d'architecture d'IA – le maillage d'IA agentique – est nécessaire pour encadrer l'évolution rapide du paysage de l'IA organisationnelle et permettre aux équipes de combiner agents sur mesure et agents prêts à l'emploi, tout en maîtrisant la dette technique croissante et les nouveaux types de risques. Mais le plus grand défi ne sera pas technique. Il sera humain : gagner la confiance, favoriser l'adoption et mettre en place une gouvernance adéquate pour gérer l'autonomie des agents et prévenir leur prolifération incontrôlée.

Pour amplifier leur impact à l'ère de l'intelligence artificielle, les organisations doivent repenser leurs approches de transformation par l'IA : passer d'initiatives dispersées à des programmes stratégiques ; des cas d'utilisation aux processus métier ; des équipes d'IA cloisonnées aux équipes de transformation transversales ; et de l'expérimentation à une mise en œuvre industrialisée et évolutive.

Les organisations devront également mettre en place les fondements nécessaires pour opérer efficacement à l'ère des agents. Elles devront développer les compétences de leurs employés, adapter leur infrastructure technologique, accélérer la valorisation des données et déployer des mécanismes de gouvernance spécifiques aux agents. Le moment est venu de clore le chapitre de l'expérimentation en IA générale — un tournant que seul le PDG peut opérer.

Chapitre 1

Le paradoxe de l'IA générationnelle : déploiement généralisé, impact minimal

Points clés

Passez à la section suivante

Partager

Près de huit entreprises sur dix ont déployé l'IA de nouvelle génération sous une forme ou une autre, mais à peu près le même pourcentage déclare qu'elle n'a eu aucun impact significatif sur ses bénéfices.1 Nous appelons cela le « paradoxe de l'IA générale ».

Le principal problème réside dans le déséquilibre entre les cas d'usage « horizontaux » et « verticaux ». Les premiers, comme les copilotes employés et les chatbots, sont largement déployés mais leurs bénéfices restent diffus, tandis que les cas d'usage verticaux, ou spécifiques à une fonction, à plus fort impact, dépassent rarement le stade pilote en raison d'obstacles techniques, organisationnels, liés aux données et culturels.

À moins que les entreprises ne s'attaquent à ces obstacles, la promesse transformationnelle de l'IA de nouvelle génération restera largement inexploitée.

L'intelligence artificielle de nouvelle génération est partout, sauf dans les comptes de résultat des entreprises.

Partager

barre latérale

À propos de QuantumBlack, l'IA de McKinsey

Avant même l'avènement de l'IA de nouvelle génération, l'intelligence artificielle s'était déjà taillé une place essentielle dans les entreprises, en alimentant des capacités avancées de prédiction, de classification et d'optimisation. Son potentiel de valeur était déjà estimé à un montant immense, entre 11 et 18 billions de dollars à l'échelle mondiale.2— principalement dans les domaines du marketing (permettant des fonctionnalités telles que le ciblage personnalisé des e-mails et la segmentation client), des ventes (qualification des prospects) et de la chaîne

d'approvisionnement (optimisation des stocks et prévision de la demande). Pourtant, l'IA restait largement l'apanage des experts. De ce fait, son adoption par l'ensemble des employés était généralement lente. De 2018 à 2022, par exemple, l'adoption de l'IA est restée relativement stable, environ 50 % des entreprises déployant cette technologie dans une seule fonction métier, selon une étude de McKinsey (Graphique 1).

Pièce justificative 1

L'intelligence artificielle de génération a globalement accéléré le déploiement de l'IA.

L'IA de nouvelle génération a étendu la portée de l'IA traditionnelle dans trois domaines révolutionnaires : la synthèse d'informations, la génération de contenu et la communication en langage humain. McKinsey estime que cette technologie pourrait générer entre 2 600 et 4 400 milliards de dollars de valeur ajoutée, en plus de celle déjà présente dans l'IA analytique traditionnelle.3

Deux ans et demi après le lancement de ChatGPT, l'IA nouvelle génération a profondément transformé la manière dont les entreprises interagissent avec l'IA. Son potentiel de transformation réside non seulement dans les nouvelles fonctionnalités qu'elle introduit, mais aussi dans sa capacité à démocratiser l'accès aux technologies d'IA avancées au sein des organisations. Cette démocratisation a entraîné une forte augmentation de la sensibilisation à l'IA et des expérimentations associées : selon la dernière enquête mondiale de McKinsey sur l'IA ,4Plus de 78 % des entreprises utilisent désormais l'IA de nouvelle génération dans au moins une fonction commerciale (contre 55 % un an auparavant).

Toutefois, cet enthousiasme ne s'est pas encore traduit par des résultats économiques concrets. Plus de 80 % des entreprises déclarent toujours que leurs initiatives en matière d'IA de nouvelle génération n'ont aucun impact significatif sur leurs bénéfices.5De plus, seulement 1 % des entreprises que nous avons interrogées considèrent leurs stratégies en matière d'IA de

génération comme matures .6On pourrait parler du « paradoxe de l'IA générationnelle » : malgré toute l'énergie, les investissements et le potentiel entourant cette technologie, son impact à grande échelle ne s'est pas encore concrétisé pour la plupart des organisations.

Au cœur du paradoxe de l'IA générationnelle se trouve un déséquilibre entre les cas d'utilisation horizontaux et verticaux.

De nombreuses organisations ont déployé des cas d'utilisation horizontaux, tels que des copilotes et des chatbots à l'échelle de l'entreprise ; près de 70 % des entreprises du classement Fortune 500, par exemple, utilisent Microsoft 365 Copilot.7Ces outils sont généralement perçus comme des leviers d'amélioration de la productivité individuelle, permettant aux employés de gagner du temps sur les tâches routinières et d'accéder à l'information et de la synthétiser plus efficacement. Toutefois, ces améliorations, bien que réelles, ont tendance à être peu visibles parmi les employés. Par conséquent, elles ne se traduisent pas facilement par des résultats concrets, que ce soit au niveau du chiffre d'affaires ou des bénéfices.

À l'inverse, les cas d'usage verticaux — ceux intégrés à des fonctions et processus métier spécifiques — ont connu un déploiement limité dans la plupart des entreprises, malgré leur potentiel plus élevé d'impact économique direct (voir graphique 2). Selon une étude de McKinsey, moins de 10 % des cas d'usage déployés dépassent le stade du projet pilote .8Même une fois pleinement déployées, ces solutions n'ont généralement pris en charge que des étapes isolées d'un processus métier et ont fonctionné de manière réactive, sur intervention humaine, plutôt que de façon proactive ou autonome. De ce fait, leur impact sur la performance de l'entreprise est resté limité.

Pièce n° 2

Dans tous les domaines fonctionnels de l'entreprise, les cas d'utilisation de l'IA de nouvelle génération se répartissent généralement en deux catégories : horizontale et verticale.

Nous nous efforçons de garantir l'accessibilité de notre site web aux personnes en situation de handicap. Pour toute information complémentaire

concernant ce contenu, n'hésitez pas à nous contacter par courriel à l'adresse suivante : McKinsey_Website_Accessibility@mckinsey.com

Comment expliquer ce déséquilibre ? D’une part, les solutions de copilotage déployées horizontalement, telles que Microsoft Copilot ou Google AI Workspace, sont accessibles et prêtes à l’emploi, et relativement faciles à mettre en œuvre. (Dans de nombreux cas, activer Microsoft Copilot se résume à ajouter une extension à un contrat Office 365 existant, sans nécessiter de refonte des flux de travail ni d’efforts majeurs de gestion du changement.) Le déploiement rapide des chatbots d’entreprise a également été motivé par des préoccupations liées à la réduction des risques. Face à l’expérimentation par les employés de modèles de langage externes de grande taille (LLM) comme ChatGPT, de nombreuses organisations ont mis en place des alternatives internes et sécurisées afin de limiter les fuites de données et de garantir la conformité aux politiques de sécurité de l’entreprise.

Gros plan sur des brins de cordes colorées s'entremêlant.

Repenser l'entreprise grâce à la technologie et à l'IA : une transformation radicale.

Mardi 21 avril, de 11 h à 11 h 30 HAE / de 17 h à 17 h 30 HEC

Rejoignez Robert Levin et Kate Smaje , associés principaux de McKinsey et co-auteurs de « Rewired : The McKinsey Playbook on How Leading Companies Win with Technology and AI », pour une conversation sur ce que font les entreprises leaders pour transformer leurs activités grâce à l'IA.

Inscrivez-vous ici

Le déploiement limité et la portée restreinte des cas d'utilisation verticaux peuvent être attribués à six facteurs principaux :

Initiatives fragmentées. Dans de nombreuses entreprises, les cas d'usage verticaux ont été identifiés par une approche ascendante et très granulaire au sein de chaque fonction. De fait, moins de 30 % des entreprises indiquent que leur PDG pilote directement leur stratégie en matière d'IA .9Cela a

entraîné une prolifération de micro-initiatives déconnectées et une dispersion des investissements en IA, avec une coordination limitée au niveau de l'entreprise.

Absence de solutions éprouvées et prêtes à l'emploi. Contrairement aux applications horizontales standard, telles que les copilotes, les cas d'usage verticaux nécessitent souvent un développement sur mesure. De ce fait, les équipes sont fréquemment contraintes de tout construire de A à Z, en utilisant des technologies émergentes et en constante évolution avec lesquelles elles ont une expérience limitée. Si de nombreuses entreprises ont investi dans des data scientists pour développer des modèles d'IA, elles manquent souvent d'ingénieurs MLOps, pourtant essentiels pour industrialiser, déployer et maintenir ces modèles en production.

Limitations technologiques des LLM. Malgré leurs capacités impressionnantes, la première génération de LLM présentait des limitations qui ont considérablement freiné leur déploiement à grande échelle. Premièrement, les LLM peuvent produire des résultats inexacts, ce qui les rend peu fiables dans les environnements où la précision et la reproductibilité sont essentielles. De plus, malgré leur puissance, les LLM sont fondamentalement passifs ; ils n’agissent que sur commande et ne peuvent pas gérer les flux de travail de manière autonome ni prendre de décisions sans intervention humaine. Les LLM ont également éprouvé des difficultés à gérer des flux de travail complexes comportant plusieurs étapes, points de décision ou logiques de branchement. Enfin, de nombreux LLM actuels disposent d’une mémoire persistante limitée, ce qui rend difficile le suivi du contexte dans le temps ou un fonctionnement cohérent lors d’interactions prolongées.

Équipes d'IA cloisonnées. Les centres d'excellence en IA ont joué un rôle crucial dans l'accélération de la sensibilisation et de l'expérimentation au sein de nombreuses organisations. Cependant, dans bien des cas, ces équipes ont fonctionné en silos, développant des modèles d'IA indépendamment des fonctions informatiques, de données ou métiers essentielles. Cette autonomie, bien qu'utile pour le prototypage rapide, a souvent rendu les solutions difficiles à déployer à grande échelle en raison d'une mauvaise intégration aux systèmes d'entreprise, de flux de données fragmentés ou d'un manque d'alignement opérationnel.

Lacunes en matière d'accessibilité et de qualité des données. Ces lacunes concernent aussi bien les données structurées que non structurées, ces dernières restant largement non réglementées dans la plupart des organisations.

Craintes culturelles et inertie organisationnelle. Dans de nombreuses organisations, le déploiement de l'IA s'est heurté à une résistance implicite de la part des équipes opérationnelles et du management intermédiaire, en raison de la crainte de perturbations, de l'incertitude quant à l'impact sur l'emploi et du manque de familiarité avec cette technologie.

Malgré son impact limité sur les résultats financiers jusqu'à présent, la première vague d'IA de génération s'est avérée loin d'être vaine. Elle a enrichi les compétences des employés, permis une expérimentation à grande échelle, accéléré la familiarisation avec l'IA dans tous les services et aidé les organisations à développer des compétences essentielles en ingénierie rapide, en évaluation des modèles et en gouvernance. Tout cela a jeté les bases d'une seconde phase plus intégrée et transformatrice : l'avènement des agents d'IA .10

Chapitre 2

Du paradoxe à la récompense : comment les agents peuvent déployer l’IA à grande échelle

Points clés

Passez à la section suivante

Partager

En automatisant les processus métier complexes, les agents permettent d'exploiter pleinement le potentiel des cas d'usage verticaux. Les entreprises visionnaires tirent déjà parti de la puissance des agents pour transformer leurs processus clés.

Pour exploiter pleinement le potentiel des agents, les entreprises doivent réinventer leur façon de travailler : modifier les flux de tâches, redéfinir les rôles humains et construire dès le départ des processus centrés sur les agents.

Pour y parvenir, il faudra un nouveau paradigme d'architecture d'IA : le maillage d'IA agentif, capable d'intégrer des agents développés sur mesure et des agents prêts à l'emploi. Mais le plus grand défi ne sera pas technique. Il sera humain : gagner la confiance pour favoriser l'adoption et établir les protocoles de gouvernance appropriés.

La percée : L'automatisation des flux de travail complexes permet d'exploiter pleinement le potentiel des cas d'utilisation verticaux.

Les langages de modélisation linguistique (LLM) ont révolutionné la manière dont les organisations interagissent avec les données, permettant la synthèse d'informations, la génération de contenu et l'interaction en langage naturel. Cependant, malgré leur puissance, les LLM sont restés fondamentalement réactifs et isolés des systèmes d'entreprise, incapables de conserver en mémoire les interactions passées ou le contexte entre les sessions ou les requêtes. Leur rôle s'est longtemps limité à l'amélioration de la productivité individuelle par le biais de tâches isolées. Les agents d'IA marquent une évolution majeure de l'IA d'entreprise, faisant passer l'IA générative de la génération réactive de contenu à une exécution autonome et orientée vers un objectif. Ces agents peuvent comprendre les objectifs, les décomposer en sous-tâches, interagir avec les humains et les systèmes, exécuter des actions et s'adapter en temps réel, le tout avec une intervention humaine minimale. Ils y parviennent en combinant les LLM avec des composants technologiques supplémentaires offrant des capacités de mémoire, de planification, d'orchestration et d'intégration.

Grâce à ces nouvelles fonctionnalités, les agents d'IA étendent le potentiel des solutions horizontales, transformant les copilotes généralistes d'outils passifs en collaborateurs proactifs. Ces derniers ne se contentent plus de répondre aux sollicitations, mais surveillent également les tableaux de bord, déclenchent des flux de travail, assurent le suivi des actions en cours et fournissent des informations pertinentes en temps réel. Toutefois, la véritable avancée réside dans le domaine vertical, où l'IA agentique permet l'automatisation de flux de travail métier complexes impliquant de multiples étapes, acteurs et systèmes – des processus qui dépassaient auparavant les capacités des outils d'IA de première génération.

Les agents apportent bien plus que de l'efficacité : ils décuplent l'agilité opérationnelle et ouvrent de nouvelles perspectives de revenus.

Côté opérations, les agents prennent en charge les tâches routinières et gourmandes en données, permettant ainsi aux humains de se concentrer sur des missions à plus forte valeur ajoutée. Mais leur rôle ne s'arrête pas là : ils transforment les processus de cinq manières :

Les agents accélèrent l'exécution en éliminant les délais entre les tâches et en permettant le traitement parallèle. Contrairement aux flux de travail traditionnels qui reposent sur des transferts séquentiels, les agents peuvent coordonner et exécuter plusieurs étapes simultanément, réduisant ainsi le temps de cycle et améliorant la réactivité.

Les agents apportent l'adaptabilité. En ingérant des données en continu, ils peuvent ajuster les flux de processus en temps réel, en réorganisant les séquences de tâches, en réattribuant les priorités ou en signalant les anomalies avant qu'elles ne provoquent des défaillances. Les flux de travail sont ainsi non seulement plus rapides, mais aussi plus intelligents.

Les agents permettent une personnalisation. En adaptant les interactions et les décisions aux profils ou comportements individuels des clients, les agents peuvent ajuster le processus de manière dynamique afin de maximiser la satisfaction et les résultats.

Les agents apportent de la flexibilité aux opérations. Étant numériques, leur capacité d'exécution peut augmenter ou diminuer en temps réel en fonction de la charge de travail, de la saisonnalité de l'activité ou des pics d'activité imprévus – une flexibilité difficile à obtenir avec des modèles de ressources humaines fixes.

Les agents contribuent également à renforcer la résilience des opérations. En surveillant les perturbations, en réacheminant les opérations et en intervenant uniquement en cas de besoin, ils assurent la continuité des processus, qu'il s'agisse de chaînes d'approvisionnement confrontées à des retards portuaires ou de flux de travail de service s'adaptant aux pannes de système.

Dans un environnement de chaîne d'approvisionnement complexe, un agent d'IA pourrait, par exemple, jouer le rôle d'une couche d'orchestration autonome pour les opérations d'approvisionnement, d'entreposage et de distribution. Connecté aux systèmes internes (tels que le système de planification de la chaîne d'approvisionnement ou le système de gestion d'entrepôt) et aux sources de données externes (comme les prévisions météorologiques, les flux de données des fournisseurs et les signaux de la demande), cet agent pourrait prévoir la demande en continu. Il pourrait ensuite identifier les risques, tels que les retards ou les perturbations, et replanifier dynamiquement les flux de transport et de stocks. En sélectionnant le mode de transport optimal en fonction du coût, du délai et de l'impact environnemental, l'agent pourrait réallouer les stocks entre les entrepôts, négocier directement avec les systèmes externes et remonter les décisions nécessitant une expertise stratégique. Résultat : une amélioration

des niveaux de service, une réduction des coûts logistiques et une diminution des émissions.

Les agents peuvent également contribuer à stimuler la croissance du chiffre d'affaires en amplifiant les sources de revenus existantes et en en débloquant de nouvelles :

Augmenter les revenus existants. Dans le e-commerce, des agents intégrés aux boutiques en ligne ou aux applications pourraient analyser proactivement le comportement des utilisateurs, le contenu de leur panier et le contexte (par exemple, la saisonnalité ou l'historique d'achats) afin de proposer des offres de vente incitative et de vente croisée en temps réel. Dans le secteur financier, ces agents pourraient aider les clients à trouver des produits financiers adaptés à leurs besoins, tels que des prêts, des assurances ou des portefeuilles d'investissement, en leur fournissant des conseils personnalisés en fonction de leur profil financier, de leur situation personnelle et de leurs habitudes de consommation.

Création de nouvelles sources de revenus. Pour les entreprises industrielles, des agents intégrés aux produits ou équipements connectés pourraient surveiller l'utilisation, détecter les seuils de performance et débloquer automatiquement des fonctionnalités ou déclencher des interventions de maintenance, permettant ainsi des modèles de rémunération à l'usage, par abonnement ou basés sur la performance. De même, les entreprises de services pourraient intégrer leur expertise interne (raisonnement juridique, interprétation fiscale et bonnes pratiques d'approvisionnement) dans des agents d'IA proposés sous forme de logiciels en tant que service (SaaS) ou d'API à leurs clients, partenaires ou PME ne disposant pas de cette expertise en interne.

En résumé, l'IA agentielle ne se contente pas d'automatiser. Elle redéfinit la manière dont les organisations fonctionnent, s'adaptent et créent de la valeur.

Ce n'est plus de la science-fiction : les entreprises visionnaires exploitent le pouvoir des agents

Les études de cas suivantes démontrent comment QuantumBlack aide les organisations à constituer des effectifs d'agents, avec des résultats qui vont bien au-delà des gains d'efficacité.

Étude de cas 1 : Comment une banque a utilisé des « usines numériques » hybrides pour moderniser ses applications existantes

Le problème : une grande banque devait moderniser son système central informatique, composé de 400 logiciels – un projet colossal dont le budget dépassait les 600 millions de dollars. De grandes équipes de développeurs ont abordé le projet en effectuant des tâches manuelles et répétitives, ce qui a engendré des difficultés de coordination entre les différents services. Elles s'appuyaient également sur une documentation et un code souvent lents et sujets aux erreurs. Si les outils d'IA de première génération ont permis d'accélérer certaines tâches, la progression est restée lente et laborieuse.

L'approche par agents : les employés humains ont été promus à des rôles de supervision, encadrant des équipes d'agents d'IA. Chaque équipe contribue à un objectif commun selon une séquence définie (voir l'illustration 3). Ces équipes documentent a posteriori l'application existante, écrivent du nouveau code, examinent le code des autres agents et intègrent ce code dans des fonctionnalités qui sont ensuite testées par d'autres agents avant la livraison du produit final. Libérés des tâches manuelles répétitives, les superviseurs humains guident chaque étape du processus, améliorant ainsi la qualité des livrables et réduisant le nombre de sprints nécessaires à l'implémentation de nouvelles fonctionnalités.

Impact : Réduction de plus de 50 % du temps et des efforts consacrés par les équipes pionnières.

Pièce justificative 3

Une grande banque a modernisé son infrastructure technologique traditionnelle avec une usine numérique hybride IA-humain.

Nous nous efforçons de garantir l'accessibilité de notre site web aux personnes en situation de handicap. Pour toute information complémentaire concernant ce contenu, n'hésitez pas à nous contacter par courriel à l'adresse suivante : McKinsey_Website_Accessibility@mckinsey.com

Étude de cas 2 : Comment un cabinet d'études a amélioré la qualité de ses données pour obtenir des informations plus approfondies sur le marché

Le problème : une société d'études de marché et de veille stratégique consacrait des ressources considérables à garantir la qualité des données, s'appuyant sur une équipe de plus de 500 personnes chargées de collecter, structurer et codifier les données, puis de générer des analyses personnalisées pour ses clients. Ce processus, réalisé manuellement, était sujet aux erreurs, dont 80 % étaient détectées par les clients eux-mêmes.

L'approche multi-agents : une solution autonome identifie les anomalies de données et explique les variations des ventes ou des parts de marché. Elle analyse les signaux internes, tels que les modifications de la nomenclature des produits, et les événements externes identifiés par des recherches web, comme les rappels de produits ou les intempéries. Les facteurs les plus influents sont synthétisés, hiérarchisés et mis à la disposition des décideurs. Grâce à la recherche avancée et au raisonnement contextuel, les agents font souvent émerger des informations qu'il serait difficile pour des analystes humains de découvrir manuellement. Bien que le système ne soit pas encore en production, il est pleinement fonctionnel et a démontré un fort potentiel pour libérer les analystes de leurs tâches stratégiques.

Impact : Gain de productivité potentiel de plus de 60 % et économies attendues de plus de 3 millions de dollars par an.

Étude de cas 3 : Comment une banque a repensé sa méthode de création de notes d'évaluation des risques de crédit

Le problème : les chargés de clientèle d'une banque de détail passaient des semaines à rédiger et à peaufiner des notes d'évaluation du risque de crédit afin de faciliter leurs décisions et de se conformer aux exigences réglementaires (voir l'annexe 4). Ce processus les obligeait à examiner et à extraire manuellement des informations provenant d'au moins dix sources de données différentes, et à élaborer un raisonnement complexe et nuancé sur des sections interdépendantes, par exemple l'évolution conjointe des prêts, des revenus et de la trésorerie.

L'approche par agents : En étroite collaboration avec les experts en risque de crédit et les chargés de clientèle de la banque, une preuve de concept a été développée afin de transformer le flux de travail des notes de crédit grâce à des agents d'IA. Ces agents assistent les chargés de clientèle en extrayant

des données, en rédigeant des sections de notes, en générant des scores de confiance pour prioriser les analyses et en suggérant des questions de suivi pertinentes. Dans ce modèle, le rôle de l'analyste évolue de la rédaction manuelle vers la supervision stratégique et la gestion des exceptions.

Impact : Une augmentation potentielle de la productivité de 20 à 60 %, dont une amélioration de 30 % du délai de traitement des crédits.

Pièce justificative 4

Une banque de détail a utilisé des agents d'IA pour réinventer le processus de création de notes d'information sur le risque de crédit.

Nous nous efforçons de garantir l'accessibilité de notre site web aux personnes en situation de handicap. Pour toute information complémentaire concernant ce contenu, n'hésitez pas à nous contacter par courriel à l'adresse suivante : McKinsey_Website_Accessibility@mckinsey.com

Pour maximiser la valeur des agents d'IA, il faut réinventer les processus.

Pour exploiter pleinement le potentiel de l'IA dans les secteurs verticaux, il ne suffit pas d'intégrer des agents aux flux de travail existants. Il est nécessaire de repenser entièrement la conception, en passant de l'automatisation des tâches au sein d'un processus existant à la réinvention du processus dans son intégralité, avec la collaboration d'humains et d'agents. En effet, lorsque des agents sont intégrés à un processus existant sans refonte, ils servent généralement d'assistants plus rapides : ils génèrent du contenu, extraient des données ou exécutent des étapes prédéfinies. Mais le processus lui-même reste séquentiel, rigide et soumis aux contraintes humaines.

Repenser un processus autour d'agents ne se limite pas à automatiser les flux de travail existants ; cela implique de repenser entièrement l'architecture du flux de tâches. Il s'agit notamment de réorganiser les étapes, de redistribuer les responsabilités entre humains et agents, et de concevoir le processus pour exploiter pleinement les atouts de l'IA agentielle : exécution parallèle réduisant considérablement les délais, adaptabilité en temps réel aux conditions changeantes, personnalisation poussée à grande échelle et capacité flexible s'ajustant instantanément à la demande.

Prenons l'exemple d'un centre d'appels clients hypothétique. Avant l'introduction d'agents IA, ce centre utilisait des outils d'IA générale pour assister le personnel d'assistance humain : recherche d'articles dans les bases de connaissances, synthèse de l'historique des tickets et aide à la rédaction des réponses. Si cette assistance a permis d'accélérer le processus et de réduire la charge cognitive, le processus lui-même restait entièrement manuel et réactif, les agents humains gérant toujours chaque étape du diagnostic, de la coordination et de la résolution. Le potentiel d'amélioration de la productivité était modeste, avec généralement une augmentation du temps de résolution et de la productivité de 5 à 10 %.

Imaginez maintenant qu'un centre d'appels intègre des agents IA tout en conservant en grande partie son flux de travail actuel : des agents sont ajoutés pour intervenir à des étapes spécifiques sans que l'acheminement, le suivi ou la résolution des demandes ne soient modifiés de bout en bout. Ces agents peuvent classer les tickets, suggérer les causes probables, proposer des solutions et même résoudre de manière autonome les problèmes fréquents et simples (comme la réinitialisation des mots de passe). Si l'impact peut être accru – avec un gain de temps estimé entre 20 et 40 % et une réduction du volume de demandes en attente de 30 à 50 % –, les difficultés de coordination et la faible adaptabilité empêchent des gains véritablement significatifs.

Mais le véritable changement s'opère au troisième niveau, lorsque le processus du centre d'appels est repensé autour de l'autonomie des agents. Dans ce modèle, les agents IA ne se contentent pas de répondre : ils détectent proactivement les problèmes clients courants (tels que les retards de livraison, les échecs de paiement ou les interruptions de service) en analysant les tendances sur tous les canaux, anticipent les besoins probables, initient automatiquement les solutions (comme les remboursements, les commandes de nouveaux articles ou la mise à jour des informations de compte) et communiquent directement avec les clients par chat ou e-mail. Les agents humains sont repositionnés comme gestionnaires d'escalade et responsables de la qualité du service, et interviennent uniquement lorsque les agents détectent une incertitude ou une exception aux schémas habituels. L'impact à ce niveau est transformateur. Cela pourrait permettre une amélioration radicale de la productivité du service client. Jusqu'à 80 % des incidents courants pourraient être résolus de manière autonome, avec une réduction du temps de résolution de 60 à 90 %

Les agents détiennent la clé de la percée, à condition que les processus soient réinventés et non simplement optimisés.

Bien sûr, tous les processus métier ne nécessitent pas une refonte complète. Une simple automatisation des tâches suffit pour les flux de travail standardisés et répétitifs, à faible variabilité – comme le traitement de la paie, l'approbation des notes de frais ou la réinitialisation des mots de passe – où les gains proviennent principalement de la réduction des interventions manuelles. En revanche, les processus complexes, transversaux, sujets aux exceptions ou étroitement liés à la performance de l'entreprise justifient souvent une refonte totale. Parmi les indicateurs clés d'une telle refonte figurent une lourdeur de coordination, des séquences rigides qui ralentissent la réactivité, des interventions humaines fréquentes pour des décisions qui pourraient être basées sur les données, et des possibilités d'adaptation dynamique ou de personnalisation. Dans ces cas, repenser le processus en s'appuyant sur la capacité de l'agent à orchestrer, s'adapter et apprendre apporte une valeur ajoutée bien supérieure à une simple accélération des flux de travail existants.

Un nouveau paradigme d'architecture d'IA — le maillage d'IA agentique — est nécessaire pour orchestrer la création de valeur à l'ère de l'agentivité.

Pour déployer leurs agents à grande échelle, les entreprises devront relever un triple défi : gérer les nouveaux risques liés aux agents IA, combiner des systèmes d’agents personnalisés et prêts à l’emploi, et rester agiles face à l’évolution rapide des technologies (tout en évitant la dépendance vis-à-vis d’un fournisseur unique).

- **Gérer une nouvelle vague de risques**. Les agents introduisent une nouvelle catégorie de risques systémiques que les architectures d'IA traditionnelles, conçues principalement pour des cas d'usage isolés et centrés sur la

gestion des actifs numériques, n'ont jamais été conçues pour gérer : autonomie incontrôlée, accès fragmenté au système, manque d'observabilité et de traçabilité, surface d'attaque croissante, prolifération et duplication des agents. Ce qui commence comme une automatisation intelligente peut rapidement se transformer en chaos opérationnel, à moins de reposer sur des fondements privilégiant le contrôle, l'évolutivité et la confiance.

- **Combiner agents personnalisés et solutions prêtes à l'emploi.** Pour exploiter pleinement le potentiel transformateur des agents d'IA, les organisations doivent aller au-delà de la simple activation d'agents intégrés à des suites logicielles. Ces agents prêts à l'emploi peuvent certes rationaliser les flux de travail routiniers, mais ils génèrent rarement un avantage stratégique. Pour tirer pleinement parti de l'IA, il est nécessaire de développer des agents sur mesure pour les processus à fort impact, tels que la résolution de bout en bout des problèmes clients, l'orchestration adaptative de la chaîne d'approvisionnement ou la prise de décision complexe. Ces agents doivent être parfaitement alignés sur la logique, les flux de données et les leviers de création de valeur de l'entreprise, ce qui les rend difficiles à reproduire et particulièrement performants.
- **Rester agile face à l'évolution rapide des technologies**. L'IA agentique est un domaine technologique émergent, et les solutions évoluent à un rythme soutenu. Les agents devront prendre en charge les flux de travail sur plusieurs systèmes et ne devront pas être intégrés de manière rigide à une plateforme spécifique. Une architecture évolutive et indépendante des fournisseurs est donc indispensable.

Ces défis ne peuvent être relevés par la simple superposition de nouveaux composants, tels que des systèmes de stockage de mémoire ou des moteurs d'orchestration, sur les architectures d'IA existantes. Si ces capacités sont nécessaires, elles ne suffisent pas. Un changement architectural fondamental

s'impose : passer d'une infrastructure statique, centrée sur les modèles logiques, à un environnement dynamique, modulaire et gouverné, conçu spécifiquement pour l'intelligence multi-agents – le maillage d'IA multi-agents.

Le maillage d'IA agentique est un paradigme architectural composable, distribué et indépendant des fournisseurs, qui permet à de multiples agents de raisonner, de collaborer et d'agir de manière autonome sur un large éventail de systèmes, d'outils et de modèles de langage — de façon sécurisée, à grande échelle et conçu pour évoluer avec la technologie. Ce paradigme repose sur cinq principes de conception qui se renforcent mutuellement :

- **Composabilité.** Tout agent, outil ou LLM peut être intégré au maillage sans modification du système**.**
- **Intelligence distribuée.** Les tâches peuvent être décomposées et résolues par des réseaux d'agents coopérants**.**
- **Découplage par couches.** Les fonctions logiques, de mémoire, d'orchestration et d'interface sont découplées afin de maximiser la modularité.
- **Neutralité vis-à-vis des fournisseurs**. Tous les composants peuvent être mis à jour ou remplacés indépendamment au gré des évolutions technologiques, évitant ainsi la dépendance à un fournisseur unique et pérennisant l'architecture. En particulier, les standards ouverts tels que le Model Context Protocol (MCP) et Agent2Agent (A2A) sont privilégiés par rapport aux protocoles propriétaires.
- **Autonomie encadrée**. Le comportement des agents est contrôlé de manière proactive via des politiques, des autorisations et des mécanismes d'escalade intégrés qui garantissent un fonctionnement sûr et transparent

Sept capacités interconnectées du maillage d'agents d'IA

Le maillage d'IA agentique sert de couche de connexion et d'orchestration, permettant aux écosystèmes d'agents intelligents à grande échelle de fonctionner de manière sûre et efficace, et d'évoluer en continu. Il permet aux entreprises de coordonner des agents sur mesure et des agents prêts à l'emploi au sein d'un cadre unifié, de faciliter la collaboration multi-agents en leur permettant de partager le contexte et de déléguer des tâches, et d'atténuer les principaux risques tels que la prolifération des agents, la dérive

d'autonomie et le manque d'observabilité, tout en préservant l'agilité nécessaire à une évolution technologique rapide (voir l'encadré « Sept capacités interconnectées du maillage d'IA agentique »).

Modèles fondamentaux pour les agents : cinq exigences

Au-delà de cette évolution architecturale, les organisations devront également revoir leurs stratégies LLM. Chaque agent personnalisé repose sur un modèle fondamental : le moteur de raisonnement qui sous-tend la perception, la prise de décision et l'interaction. À l'ère des agents, les exigences imposées aux LLM évoluent considérablement. Les agents ne sont plus de simples copilotes passifs ; ce sont des systèmes autonomes, persistants et embarqués. Il en résulte cinq catégories critiques d'exigences LLM, chacune associée à des contextes de déploiement spécifiques, pour lesquels différents types de modèles seront pertinents (voir l'encadré « Modèles fondamentaux pour agents : cinq exigences »).

Enfin, pour déployer véritablement des agents à l'échelle de l'entreprise, les systèmes d'entreprise eux-mêmes doivent également évoluer.

À court terme, les API (protocoles permettant à différentes applications logicielles de communiquer et d'échanger des données) resteront l'interface principale d'interaction entre les agents et les systèmes d'entreprise. Mais à long terme, les API seules ne suffiront plus. Les organisations doivent repenser leurs architectures informatiques autour d'un modèle centré sur l'agent, où les interfaces utilisateur, la logique et les couches d'accès aux données sont conçues nativement pour l'interaction machine plutôt que pour la navigation humaine. Dans un tel modèle, les systèmes ne sont plus organisés autour d'écrans et de formulaires, mais autour d'interfaces lisibles par machine, de flux de travail autonomes et de processus de décision pilotés par les agents.

Cette transformation est déjà en cours. Microsoft intègre des agents au cœur de Dynamics 365 et de Microsoft 365 via Copilot Studio ; Salesforce étend Agentforce en une couche d'orchestration multi-agents ; SAP repense sa plateforme technologique d'entreprise (BTP) pour prendre en charge l'intégration d'agents via Joule. Ces changements annoncent une transition plus large : l'avenir des logiciels d'entreprise ne repose pas uniquement sur l'IA, mais sur une approche nativement axée sur les agents.

Le principal défi ne sera pas technique, il sera humain

À mesure que les agents évoluent de copilotes passifs à acteurs proactifs – et se déploient à l'échelle de l'entreprise – la complexité qu'ils introduisent sera non seulement technique, mais surtout organisationnelle. Le véritable défi réside dans la coordination, le jugement et la confiance. Cette complexité organisationnelle se manifestera principalement selon trois axes : la manière dont humains et agents

cohabitent au quotidien ; la façon dont les organisations mettent en place une gouvernance pour les systèmes autonomes ; et la manière dont elles préviennent la prolifération incontrôlée des agents à mesure que leur création se démocratise**.**

- **Cohabitation humain-agent.** Les agents ne se contenteront pas d'assister les humains ; ils agiront à leurs côtés. Ceci soulève des questions complexes sur l'interaction et la coexistence : quand un agent doit-il prendre l'initiative ? Quand doit-il s'en remettre à l'humain ? Comment préserver l'autonomie et le contrôle humains sans compromettre les avantages mêmes qu'apportent les agents ? Clarifier ces rôles nécessitera du temps, des expérimentations et une adaptation culturelle. La confiance ne reposera pas uniquement sur les performances techniques ; elle dépendra de la transparence de la communication des agents, de la prévisibilité de leur comportement et de leur intégration intuitive aux flux de travail quotidiens.
-
- Saisir l'avantage de l'IA agentive
- 13 juin 2025| Rapport
-
-
- Share
- Print
-
- Download
- Sauvegarder
- Un guide pratique pour les PDG afin de résoudre le paradoxe de l'IA générationnelle et de débloquer un impact à grande échelle grâce aux agents d'IA.
- TÉLÉCHARGEMENTS
- Rapport complet (28 pages)
- En un coup d'œil
- Chapitre 1
- Chapitre 2
-
- Chapitre 3
- Conclusion
- En un coup d'œil
- Près de huit entreprises sur dix déclarent utiliser l'IA de nouvelle génération, mais tout autant ne constatent aucun impact significatif sur leurs résultats financiers.1Considérez cela comme le « paradoxe de l'IA générationnelle ».
-
- Signature

- À propos des auteurs
- Au cœur de ce paradoxe se trouve un déséquilibre entre les copilotes et les chatbots « horizontaux » (à l'échelle de l'entreprise) — qui se sont développés rapidement mais n'offrent que des gains diffus et difficiles à mesurer — et les cas d'utilisation « verticaux » (spécifiques à une fonction) plus transformateurs — dont environ 90 % restent bloqués en mode pilote.
- Les agents d'IA offrent une solution au paradoxe de l'IA généraliste. En effet, ils ont le potentiel d'automatiser des processus métier complexes — en combinant autonomie, planification, mémoire et intégration — transformant ainsi l'IA généraliste d'un outil réactif en un collaborateur virtuel proactif et orienté vers un objectif.
- Ce changement permet bien plus qu'une simple amélioration de l'efficacité. Les agents décuplent l'agilité opérationnelle et créent de nouvelles sources de revenus.
- Mais pour exploiter pleinement le potentiel de l'IA agentielle, il ne suffit pas d'intégrer des agents aux flux de travail existants. Il faut repenser ces flux de travail de A à Z, en plaçant les agents au cœur du système.
-
- Partager
-
- barre latérale
- Avant-propos
- Un nouveau paradigme d'architecture d'IA – le maillage d'IA agentique – est nécessaire pour encadrer l'évolution rapide du paysage de l'IA organisationnelle et permettre aux équipes de combiner agents sur mesure et agents prêts à l'emploi, tout en maîtrisant la dette technique croissante et les nouveaux types de risques. Mais le plus grand défi ne sera pas technique. Il sera humain : gagner la confiance, favoriser l'adoption et mettre en place une gouvernance adéquate pour gérer l'autonomie des agents et prévenir leur prolifération incontrôlée.
- Pour amplifier leur impact à l'ère de l'intelligence artificielle, les organisations doivent repenser leurs approches de transformation par l'IA : passer d'initiatives dispersées à des programmes stratégiques ; des cas d'utilisation aux processus métier ; des équipes d'IA cloisonnées aux équipes de transformation transversales ; et de l'expérimentation à une mise en œuvre industrialisée et évolutive.
- Les organisations devront également mettre en place les fondements nécessaires pour opérer efficacement à l'ère des agents. Elles devront développer les compétences de leurs employés, adapter leur infrastructure technologique, accélérer la valorisation des données et déployer des mécanismes de gouvernance spécifiques aux agents. Le moment est venu

de clore le chapitre de l'expérimentation en IA générale — un tournant que seul le PDG peut opérer.

- Chapitre 1
- Le paradoxe de l'IA générationnelle : déploiement généralisé, impact minimal
- Points clés
- Passez à la section suivante
-
-
- Partager
- Près de huit entreprises sur dix ont déployé l'IA de nouvelle génération sous une forme ou une autre, mais à peu près le même pourcentage déclare qu'elle n'a eu aucun impact significatif sur ses bénéfices.1 Nous appelons cela le « paradoxe de l'IA générale ».
- Le principal problème réside dans le déséquilibre entre les cas d'usage « horizontaux » et « verticaux ». Les premiers, comme les copilotes employés et les chatbots, sont largement déployés mais leurs bénéfices restent diffus, tandis que les cas d'usage verticaux, ou spécifiques à une fonction, à plus fort impact, dépassent rarement le stade pilote en raison d'obstacles techniques, organisationnels, liés aux données et culturels.
- À moins que les entreprises ne s'attaquent à ces obstacles, la promesse transformationnelle de l'IA de nouvelle génération restera largement inexploitée.
- L'intelligence artificielle de nouvelle génération est partout, sauf dans les comptes de résultat des entreprises.
-
- Partager
-
- barre latérale
- À propos de QuantumBlack, l'IA de McKinsey
- Avant même l'avènement de l'IA de nouvelle génération, l'intelligence artificielle s'était déjà taillé une place essentielle dans les entreprises, en alimentant des capacités avancées de prédiction, de classification et d'optimisation. Son potentiel de valeur était déjà estimé à un montant immense, entre 11 et 18 billions de dollars à l'échelle mondiale.2— principalement dans les domaines du marketing (permettant des fonctionnalités telles que le ciblage personnalisé des e-mails et la segmentation client), des ventes (qualification des prospects) et de la chaîne d'approvisionnement (optimisation des stocks et prévision de la demande). Pourtant, l'IA restait largement l'apanage des experts. De ce fait, son adoption par l'ensemble des employés était généralement lente. De 2018 à 2022, par exemple, l'adoption de l'IA est restée relativement stable,

environ 50 % des entreprises déployant cette technologie dans une seule fonction métier, selon une étude de McKinsey (Graphique 1).

-
- Pièce justificative 1
- L'intelligence artificielle de génération a globalement accéléré le déploiement de l'IA.
- Nous nous efforçons de garantir l'accessibilité de notre site web aux personnes en situation de handicap. Pour toute information complémentaire concernant ce contenu, n'hésitez pas à nous contacter par courriel à l'adresse suivante : McKinsey_Website_Accessibility@mckinsey.com
- L'IA de nouvelle génération a étendu la portée de l'IA traditionnelle dans trois domaines révolutionnaires : la synthèse d'informations, la génération de contenu et la communication en langage humain. McKinsey estime que cette technologie pourrait générer entre 2 600 et 4 400 milliards de dollars de valeur ajoutée, en plus de celle déjà présente dans l'IA analytique traditionnelle.3
-
- Deux ans et demi après le lancement de ChatGPT, l'IA nouvelle génération a profondément transformé la manière dont les entreprises interagissent avec l'IA. Son potentiel de transformation réside non seulement dans les nouvelles fonctionnalités qu'elle introduit, mais aussi dans sa capacité à démocratiser l'accès aux technologies d'IA avancées au sein des organisations. Cette démocratisation a entraîné une forte augmentation de la sensibilisation à l'IA et des expérimentations associées : selon la dernière enquête mondiale de McKinsey sur l'IA ,4Plus de 78 % des entreprises utilisent désormais l'IA de nouvelle génération dans au moins une fonction commerciale (contre 55 % un an auparavant).
-
- Toutefois, cet enthousiasme ne s'est pas encore traduit par des résultats économiques concrets. Plus de 80 % des entreprises déclarent toujours que leurs initiatives en matière d'IA de nouvelle génération n'ont aucun impact significatif sur leurs bénéfices.5De plus, seulement 1 % des entreprises que nous avons interrogées considèrent leurs stratégies en matière d'IA de génération comme matures .6On pourrait parler du « paradoxe de l’IA générationnelle » : malgré toute l’énergie, les investissements et le potentiel entourant cette technologie, son impact à grande échelle ne s’est pas encore concrétisé pour la plupart des organisations.
-
- Au cœur du paradoxe de l'IA générationnelle se trouve un déséquilibre entre les cas d'utilisation horizontaux et verticaux.
- De nombreuses organisations ont déployé des cas d'utilisation horizontaux, tels que des copilotes et des chatbots à l'échelle de l'entreprise ; près de 70

% des entreprises du classement Fortune 500, par exemple, utilisent Microsoft 365 Copilot.7Ces outils sont généralement perçus comme des leviers d'amélioration de la productivité individuelle, permettant aux employés de gagner du temps sur les tâches routinières et d'accéder à l'information et de la synthétiser plus efficacement. Toutefois, ces améliorations, bien que réelles, ont tendance à être peu visibles parmi les employés. Par conséquent, elles ne se traduisent pas facilement par des résultats concrets, que ce soit au niveau du chiffre d'affaires ou des bénéfices.

-
- À l’inverse, les cas d’usage verticaux — ceux intégrés à des fonctions et processus métier spécifiques — ont connu un déploiement limité dans la plupart des entreprises, malgré leur potentiel plus élevé d’impact économique direct (voir graphique 2). Selon une étude de McKinsey, moins de 10 % des cas d’usage déployés dépassent le stade du projet pilote .8Même une fois pleinement déployées, ces solutions n'ont généralement pris en charge que des étapes isolées d'un processus métier et ont fonctionné de manière réactive, sur intervention humaine, plutôt que de façon proactive ou autonome. De ce fait, leur impact sur la performance de l'entreprise est resté limité.
-
- Pièce n° 2
- Dans tous les domaines fonctionnels de l'entreprise, les cas d'utilisation de l'IA de nouvelle génération se répartissent généralement en deux catégories : horizontale et verticale.
- Nous nous efforçons de garantir l'accessibilité de notre site web aux personnes en situation de handicap. Pour toute information complémentaire concernant ce contenu, n'hésitez pas à nous contacter par courriel à l'adresse suivante : McKinsey_Website_Accessibility@mckinsey.com
- Comment expliquer ce déséquilibre ? D’une part, les solutions de copilotage déployées horizontalement, telles que Microsoft Copilot ou Google AI Workspace, sont accessibles et prêtes à l’emploi, et relativement faciles à mettre en œuvre. (Dans de nombreux cas, activer Microsoft Copilot se résume à ajouter une extension à un contrat Office 365 existant, sans nécessiter de refonte des flux de travail ni d’efforts majeurs de gestion du changement.) Le déploiement rapide des chatbots d’entreprise a également été motivé par des préoccupations liées à la réduction des risques. Face à l’expérimentation par les employés de modèles de langage externes de grande taille (LLM) comme ChatGPT, de nombreuses organisations ont mis en place des alternatives internes et sécurisées afin de limiter les fuites de données et de garantir la conformité aux politiques de sécurité de l’entreprise.

-
- Gros plan sur des brins de cordes colorées s'entremêlant.
- Repenser l'entreprise grâce à la technologie et à l'IA : une transformation radicale.
- Mardi 21 avril, de 11 h à 11 h 30 HAE / de 17 h à 17 h 30 HEC
-
- Rejoignez Robert Levin et Kate Smaje , associés principaux de McKinsey et co-auteurs de « Rewired : The McKinsey Playbook on How Leading Companies Win with Technology and AI », pour une conversation sur ce que font les entreprises leaders pour transformer leurs activités grâce à l'IA.
-
-
- Inscrivez-vous ici
- Le déploiement limité et la portée restreinte des cas d'utilisation verticaux peuvent être attribués à six facteurs principaux :
-
- Initiatives fragmentées. Dans de nombreuses entreprises, les cas d'usage verticaux ont été identifiés par une approche ascendante et très granulaire au sein de chaque fonction. De fait, moins de 30 % des entreprises indiquent que leur PDG pilote directement leur stratégie en matière d'IA .9Cela a entraîné une prolifération de micro-initiatives déconnectées et une dispersion des investissements en IA, avec une coordination limitée au niveau de l'entreprise.
- Absence de solutions éprouvées et prêtes à l'emploi. Contrairement aux applications horizontales standard, telles que les copilotes, les cas d'usage verticaux nécessitent souvent un développement sur mesure. De ce fait, les équipes sont fréquemment contraintes de tout construire de A à Z, en utilisant des technologies émergentes et en constante évolution avec lesquelles elles ont une expérience limitée. Si de nombreuses entreprises ont investi dans des data scientists pour développer des modèles d'IA, elles manquent souvent d'ingénieurs MLOps, pourtant essentiels pour industrialiser, déployer et maintenir ces modèles en production.
- Limitations technologiques des LLM. Malgré leurs capacités impressionnantes, la première génération de LLM présentait des limitations qui ont considérablement freiné leur déploiement à grande échelle. Premièrement, les LLM peuvent produire des résultats inexacts, ce qui les rend peu fiables dans les environnements où la précision et la reproductibilité sont essentielles. De plus, malgré leur puissance, les LLM sont fondamentalement passifs ; ils n’agissent que sur commande et ne peuvent pas gérer les flux de travail de manière autonome ni prendre de décisions sans intervention humaine. Les LLM ont également éprouvé des difficultés à gérer des flux de travail complexes comportant plusieurs

étapes, points de décision ou logiques de branchement. Enfin, de nombreux LLM actuels disposent d'une mémoire persistante limitée, ce qui rend difficile le suivi du contexte dans le temps ou un fonctionnement cohérent lors d'interactions prolongées.

- Équipes d'IA cloisonnées. Les centres d'excellence en IA ont joué un rôle crucial dans l'accélération de la sensibilisation et de l'expérimentation au sein de nombreuses organisations. Cependant, dans bien des cas, ces équipes ont fonctionné en silos, développant des modèles d'IA indépendamment des fonctions informatiques, de données ou métiers essentielles. Cette autonomie, bien qu'utile pour le prototypage rapide, a souvent rendu les solutions difficiles à déployer à grande échelle en raison d'une mauvaise intégration aux systèmes d'entreprise, de flux de données fragmentés ou d'un manque d'alignement opérationnel.
- Lacunes en matière d'accessibilité et de qualité des données. Ces lacunes concernent aussi bien les données structurées que non structurées, ces dernières restant largement non réglementées dans la plupart des organisations.
- Craintes culturelles et inertie organisationnelle. Dans de nombreuses organisations, le déploiement de l'IA s'est heurté à une résistance implicite de la part des équipes opérationnelles et du management intermédiaire, en raison de la crainte de perturbations, de l'incertitude quant à l'impact sur l'emploi et du manque de familiarité avec cette technologie.
- Malgré son impact limité sur les résultats financiers jusqu'à présent, la première vague d'IA de génération s'est avérée loin d'être vaine. Elle a enrichi les compétences des employés, permis une expérimentation à grande échelle, accéléré la familiarisation avec l'IA dans tous les services et aidé les organisations à développer des compétences essentielles en ingénierie rapide, en évaluation des modèles et en gouvernance. Tout cela a jeté les bases d'une seconde phase plus intégrée et transformatrice : l'avènement des agents d'IA .10
-
- Chapitre 2
- Du paradoxe à la récompense : comment les agents peuvent déployer l'IA à grande échelle
- Points clés
- Passez à la section suivante
-
-
- Partager
- En automatisant les processus métier complexes, les agents permettent d'exploiter pleinement le potentiel des cas d'usage verticaux. Les entreprises visionnaires tirent déjà parti de la puissance des agents pour transformer leurs processus clés.

- Pour exploiter pleinement le potentiel des agents, les entreprises doivent réinventer leur façon de travailler : modifier les flux de tâches, redéfinir les rôles humains et construire dès le départ des processus centrés sur les agents.
- Pour y parvenir, il faudra un nouveau paradigme d'architecture d'IA : le maillage d'IA agentif, capable d'intégrer des agents développés sur mesure et des agents prêts à l'emploi. Mais le plus grand défi ne sera pas technique. Il sera humain : gagner la confiance pour favoriser l'adoption et établir les protocoles de gouvernance appropriés.
- La percée : L'automatisation des flux de travail complexes permet d'exploiter pleinement le potentiel des cas d'utilisation verticaux.
- Les langages de modélisation linguistique (LLM) ont révolutionné la manière dont les organisations interagissent avec les données, permettant la synthèse d'informations, la génération de contenu et l'interaction en langage naturel. Cependant, malgré leur puissance, les LLM sont restés fondamentalement réactifs et isolés des systèmes d'entreprise, incapables de conserver en mémoire les interactions passées ou le contexte entre les sessions ou les requêtes. Leur rôle s'est longtemps limité à l'amélioration de la productivité individuelle par le biais de tâches isolées. Les agents d'IA marquent une évolution majeure de l'IA d'entreprise, faisant passer l'IA générative de la génération réactive de contenu à une exécution autonome et orientée vers un objectif. Ces agents peuvent comprendre les objectifs, les décomposer en sous-tâches, interagir avec les humains et les systèmes, exécuter des actions et s'adapter en temps réel, le tout avec une intervention humaine minimale. Ils y parviennent en combinant les LLM avec des composants technologiques supplémentaires offrant des capacités de mémoire, de planification, d'orchestration et d'intégration.
-
- Grâce à ces nouvelles fonctionnalités, les agents d'IA étendent le potentiel des solutions horizontales, transformant les copilotes généralistes d'outils passifs en collaborateurs proactifs. Ces derniers ne se contentent plus de répondre aux sollicitations, mais surveillent également les tableaux de bord, déclenchent des flux de travail, assurent le suivi des actions en cours et fournissent des informations pertinentes en temps réel. Toutefois, la véritable avancée réside dans le domaine vertical, où l'IA agentique permet l'automatisation de flux de travail métier complexes impliquant de multiples étapes, acteurs et systèmes – des processus qui dépassaient auparavant les capacités des outils d'IA de première génération.
-
- Les agents apportent bien plus que de l'efficacité : ils décuplent l'agilité opérationnelle et ouvrent de nouvelles perspectives de revenus.
- Côté opérations, les agents prennent en charge les tâches routinières et gourmandes en données, permettant ainsi aux humains de se concentrer sur

des missions à plus forte valeur ajoutée. Mais leur rôle ne s'arrête pas là : ils transforment les processus de cinq manières :

-
- Les agents accélèrent l'exécution en éliminant les délais entre les tâches et en permettant le traitement parallèle. Contrairement aux flux de travail traditionnels qui reposent sur des transferts séquentiels, les agents peuvent coordonner et exécuter plusieurs étapes simultanément, réduisant ainsi le temps de cycle et améliorant la réactivité.
- Les agents apportent l'adaptabilité. En ingérant des données en continu, ils peuvent ajuster les flux de processus en temps réel, en réorganisant les séquences de tâches, en réattribuant les priorités ou en signalant les anomalies avant qu'elles ne provoquent des défaillances. Les flux de travail sont ainsi non seulement plus rapides, mais aussi plus intelligents.
- Les agents permettent une personnalisation. En adaptant les interactions et les décisions aux profils ou comportements individuels des clients, les agents peuvent ajuster le processus de manière dynamique afin de maximiser la satisfaction et les résultats.
- Les agents apportent de la flexibilité aux opérations. Étant numériques, leur capacité d'exécution peut augmenter ou diminuer en temps réel en fonction de la charge de travail, de la saisonnalité de l'activité ou des pics d'activité imprévus – une flexibilité difficile à obtenir avec des modèles de ressources humaines fixes.
- Les agents contribuent également à renforcer la résilience des opérations. En surveillant les perturbations, en réacheminant les opérations et en intervenant uniquement en cas de besoin, ils assurent la continuité des processus, qu'il s'agisse de chaînes d'approvisionnement confrontées à des retards portuaires ou de flux de travail de service s'adaptant aux pannes de système.
- Dans un environnement de chaîne d'approvisionnement complexe, un agent d'IA pourrait, par exemple, jouer le rôle d'une couche d'orchestration autonome pour les opérations d'approvisionnement, d'entreposage et de distribution. Connecté aux systèmes internes (tels que le système de planification de la chaîne d'approvisionnement ou le système de gestion d'entrepôt) et aux sources de données externes (comme les prévisions météorologiques, les flux de données des fournisseurs et les signaux de la demande), cet agent pourrait prévoir la demande en continu. Il pourrait ensuite identifier les risques, tels que les retards ou les perturbations, et replanifier dynamiquement les flux de transport et de stocks. En sélectionnant le mode de transport optimal en fonction du coût, du délai et de l'impact environnemental, l'agent pourrait réallouer les stocks entre les entrepôts, négocier directement avec les systèmes externes et remonter les décisions nécessitant une expertise stratégique. Résultat : une amélioration

des niveaux de service, une réduction des coûts logistiques et une diminution des émissions.

-
- Les agents peuvent également contribuer à stimuler la croissance du chiffre d'affaires en amplifiant les sources de revenus existantes et en en débloquant de nouvelles :
-
- Augmenter les revenus existants. Dans le e-commerce, des agents intégrés aux boutiques en ligne ou aux applications pourraient analyser proactivement le comportement des utilisateurs, le contenu de leur panier et le contexte (par exemple, la saisonnalité ou l'historique d'achats) afin de proposer des offres de vente incitative et de vente croisée en temps réel. Dans le secteur financier, ces agents pourraient aider les clients à trouver des produits financiers adaptés à leurs besoins, tels que des prêts, des assurances ou des portefeuilles d'investissement, en leur fournissant des conseils personnalisés en fonction de leur profil financier, de leur situation personnelle et de leurs habitudes de consommation.
- Création de nouvelles sources de revenus. Pour les entreprises industrielles, des agents intégrés aux produits ou équipements connectés pourraient surveiller l'utilisation, détecter les seuils de performance et débloquer automatiquement des fonctionnalités ou déclencher des interventions de maintenance, permettant ainsi des modèles de rémunération à l'usage, par abonnement ou basés sur la performance. De même, les entreprises de services pourraient intégrer leur expertise interne (raisonnement juridique, interprétation fiscale et bonnes pratiques d'approvisionnement) dans des agents d'IA proposés sous forme de logiciels en tant que service (SaaS) ou d'API à leurs clients, partenaires ou PME ne disposant pas de cette expertise en interne.
- En résumé, l'IA agentielle ne se contente pas d'automatiser. Elle redéfinit la manière dont les organisations fonctionnent, s'adaptent et créent de la valeur.
-
- Ce n'est plus de la science-fiction : les entreprises visionnaires exploitent le pouvoir des agents
- Les études de cas suivantes démontrent comment QuantumBlack aide les organisations à constituer des effectifs d'agents, avec des résultats qui vont bien au-delà des gains d'efficacité.
-
- Étude de cas 1 : Comment une banque a utilisé des « usines numériques » hybrides pour moderniser ses applications existantes
- Le problème : une grande banque devait moderniser son système central informatique, composé de 400 logiciels – un projet colossal dont le budget dépassait les 600 millions de dollars. De grandes équipes de développeurs

ont abordé le projet en effectuant des tâches manuelles et répétitives, ce qui a engendré des difficultés de coordination entre les différents services. Elles s'appuyaient également sur une documentation et un code souvent lents et sujets aux erreurs. Si les outils d'IA de première génération ont permis d'accélérer certaines tâches, la progression est restée lente et laborieuse.

-
- L'approche par agents : les employés humains ont été promus à des rôles de supervision, encadrant des équipes d'agents d'IA. Chaque équipe contribue à un objectif commun selon une séquence définie (voir l'illustration 3). Ces équipes documentent a posteriori l'application existante, écrivent du nouveau code, examinent le code des autres agents et intègrent ce code dans des fonctionnalités qui sont ensuite testées par d'autres agents avant la livraison du produit final. Libérés des tâches manuelles répétitives, les superviseurs humains guident chaque étape du processus, améliorant ainsi la qualité des livrables et réduisant le nombre de sprints nécessaires à l'implémentation de nouvelles fonctionnalités.
-
- Impact : Réduction de plus de 50 % du temps et des efforts consacrés par les équipes pionnières.
-
- Pièce justificative 3
- Une grande banque a modernisé son infrastructure technologique traditionnelle avec une usine numérique hybride IA-humain.
- Nous nous efforçons de garantir l'accessibilité de notre site web aux personnes en situation de handicap. Pour toute information complémentaire concernant ce contenu, n'hésitez pas à nous contacter par courriel à l'adresse suivante : McKinsey_Website_Accessibility@mckinsey.com
- Étude de cas 2 : Comment un cabinet d'études a amélioré la qualité de ses données pour obtenir des informations plus approfondies sur le marché
- Le problème : une société d'études de marché et de veille stratégique consacrait des ressources considérables à garantir la qualité des données, s'appuyant sur une équipe de plus de 500 personnes chargées de collecter, structurer et codifier les données, puis de générer des analyses personnalisées pour ses clients. Ce processus, réalisé manuellement, était sujet aux erreurs, dont 80 % étaient détectées par les clients eux-mêmes.
-
- L'approche multi-agents : une solution autonome identifie les anomalies de données et explique les variations des ventes ou des parts de marché. Elle analyse les signaux internes, tels que les modifications de la nomenclature des produits, et les événements externes identifiés par des recherches web, comme les rappels de produits ou les intempéries. Les facteurs les plus

influents sont synthétisés, hiérarchisés et mis à la disposition des décideurs. Grâce à la recherche avancée et au raisonnement contextuel, les agents font souvent émerger des informations qu'il serait difficile pour des analystes humains de découvrir manuellement. Bien que le système ne soit pas encore en production, il est pleinement fonctionnel et a démontré un fort potentiel pour libérer les analystes de leurs tâches stratégiques.

-
- Impact : Gain de productivité potentiel de plus de 60 % et économies attendues de plus de 3 millions de dollars par an.
-
- Étude de cas 3 : Comment une banque a repensé sa méthode de création de notes d’évaluation des risques de crédit
- Le problème : les chargés de clientèle d’une banque de détail passaient des semaines à rédiger et à peaufiner des notes d’évaluation du risque de crédit afin de faciliter leurs décisions et de se conformer aux exigences réglementaires (voir l’annexe 4). Ce processus les obligeait à examiner et à extraire manuellement des informations provenant d’au moins dix sources de données différentes, et à élaborer un raisonnement complexe et nuancé sur des sections interdépendantes, par exemple l’évolution conjointe des prêts, des revenus et de la trésorerie.
-
- L’approche par agents : En étroite collaboration avec les experts en risque de crédit et les chargés de clientèle de la banque, une preuve de concept a été développée afin de transformer le flux de travail des notes de crédit grâce à des agents d’IA. Ces agents assistent les chargés de clientèle en extrayant des données, en rédigeant des sections de notes, en générant des scores de confiance pour prioriser les analyses et en suggérant des questions de suivi pertinentes. Dans ce modèle, le rôle de l’analyste évolue de la rédaction manuelle vers la supervision stratégique et la gestion des exceptions.
-
- Impact : Une augmentation potentielle de la productivité de 20 à 60 %, dont une amélioration de 30 % du délai de traitement des crédits.
-
- Pièce justificative 4
- Une banque de détail a utilisé des agents d'IA pour réinventer le processus de création de notes d'information sur le risque de crédit.
- Nous nous efforçons de garantir l'accessibilité de notre site web aux personnes en situation de handicap. Pour toute information complémentaire concernant ce contenu, n'hésitez pas à nous contacter par courriel à l'adresse suivante : McKinsey_Website_Accessibility@mckinsey.com
- Pour maximiser la valeur des agents d'IA, il faut réinventer les processus.

- Pour exploiter pleinement le potentiel de l'IA dans les secteurs verticaux, il ne suffit pas d'intégrer des agents aux flux de travail existants. Il est nécessaire de repenser entièrement la conception, en passant de l'automatisation des tâches au sein d'un processus existant à la réinvention du processus dans son intégralité, avec la collaboration d'humains et d'agents. En effet, lorsque des agents sont intégrés à un processus existant sans refonte, ils servent généralement d'assistants plus rapides : ils génèrent du contenu, extraient des données ou exécutent des étapes prédéfinies. Mais le processus lui-même reste séquentiel, rigide et soumis aux contraintes humaines.
-
- Repenser un processus autour d'agents ne se limite pas à automatiser les flux de travail existants ; cela implique de repenser entièrement l'architecture du flux de tâches. Il s'agit notamment de réorganiser les étapes, de redistribuer les responsabilités entre humains et agents, et de concevoir le processus pour exploiter pleinement les atouts de l'IA agentielle : exécution parallèle réduisant considérablement les délais, adaptabilité en temps réel aux conditions changeantes, personnalisation poussée à grande échelle et capacité flexible s'ajustant instantanément à la demande.
-
- Prenons l'exemple d'un centre d'appels clients hypothétique. Avant l'introduction d'agents IA, ce centre utilisait des outils d'IA générale pour assister le personnel d'assistance humain : recherche d'articles dans les bases de connaissances, synthèse de l'historique des tickets et aide à la rédaction des réponses. Si cette assistance a permis d'accélérer le processus et de réduire la charge cognitive, le processus lui-même restait entièrement manuel et réactif, les agents humains gérant toujours chaque étape du diagnostic, de la coordination et de la résolution. Le potentiel d'amélioration de la productivité était modeste, avec généralement une augmentation du temps de résolution et de la productivité de 5 à 10 %.
-
- Imaginez maintenant qu'un centre d'appels intègre des agents IA tout en conservant en grande partie son flux de travail actuel : des agents sont ajoutés pour intervenir à des étapes spécifiques sans que l'acheminement, le suivi ou la résolution des demandes ne soient modifiés de bout en bout. Ces agents peuvent classer les tickets, suggérer les causes probables, proposer des solutions et même résoudre de manière autonome les problèmes fréquents et simples (comme la réinitialisation des mots de passe). Si l'impact peut être accru – avec un gain de temps estimé entre 20 et 40 % et une réduction du volume de demandes en attente de 30 à 50 % – , les difficultés de coordination et la faible adaptabilité empêchent des gains véritablement significatifs.

-
- Mais le véritable changement s'opère au troisième niveau, lorsque le processus du centre d'appels est repensé autour de l'autonomie des agents. Dans ce modèle, les agents IA ne se contentent pas de répondre : ils détectent proactivement les problèmes clients courants (tels que les retards de livraison, les échecs de paiement ou les interruptions de service) en analysant les tendances sur tous les canaux, anticipent les besoins probables, initient automatiquement les solutions (comme les remboursements, les commandes de nouveaux articles ou la mise à jour des informations de compte) et communiquent directement avec les clients par chat ou e-mail. Les agents humains sont repositionnés comme gestionnaires d'escalade et responsables de la qualité du service, et interviennent uniquement lorsque les agents détectent une incertitude ou une exception aux schémas habituels. L'impact à ce niveau est transformateur. Cela pourrait permettre une amélioration radicale de la productivité du service client. Jusqu'à 80 % des incidents courants pourraient être résolus de manière autonome, avec une réduction du temps de résolution de 60 à 90 % (voir l'illustration 5).
-
- Pièce à conviction 5
- Les agents détiennent la clé de la percée, à condition que les processus soient réinventés et non simplement optimisés.
- Nous nous efforçons de garantir l'accessibilité de notre site web aux personnes en situation de handicap. Pour toute information complémentaire concernant ce contenu, n'hésitez pas à nous contacter par courriel à l'adresse suivante : McKinsey_Website_Accessibility@mckinsey.com
- Bien sûr, tous les processus métier ne nécessitent pas une refonte complète. Une simple automatisation des tâches suffit pour les flux de travail standardisés et répétitifs, à faible variabilité – comme le traitement de la paie, l'approbation des notes de frais ou la réinitialisation des mots de passe – où les gains proviennent principalement de la réduction des interventions manuelles. En revanche, les processus complexes, transversaux, sujets aux exceptions ou étroitement liés à la performance de l'entreprise justifient souvent une refonte totale. Parmi les indicateurs clés d'une telle refonte figurent une lourdeur de coordination, des séquences rigides qui ralentissent la réactivité, des interventions humaines fréquentes pour des décisions qui pourraient être basées sur les données, et des possibilités d'adaptation dynamique ou de personnalisation. Dans ces cas, repenser le processus en s'appuyant sur la capacité de l'agent à orchestrer, s'adapter et apprendre apporte une valeur ajoutée bien supérieure à une simple accélération des flux de travail existants.
-

- Un nouveau paradigme d'architecture d'IA — le maillage d'IA agentique — est nécessaire pour orchestrer la création de valeur à l'ère de l'agentivité.
- Pour déployer leurs agents à grande échelle, les entreprises devront relever un triple défi : gérer les nouveaux risques liés aux agents IA, combiner des systèmes d'agents personnalisés et prêts à l'emploi, et rester agiles face à l'évolution rapide des technologies (tout en évitant la dépendance vis-à-vis d'un fournisseur unique).
-
- Gérer une nouvelle vague de risques. Les agents introduisent une nouvelle catégorie de risques systémiques que les architectures d'IA traditionnelles, conçues principalement pour des cas d'usage isolés et centrés sur la gestion des actifs numériques, n'ont jamais été conçues pour gérer : autonomie incontrôlée, accès fragmenté au système, manque d'observabilité et de traçabilité, surface d'attaque croissante, prolifération et duplication des agents. Ce qui commence comme une automatisation intelligente peut rapidement se transformer en chaos opérationnel, à moins de reposer sur des fondements privilégiant le contrôle, l'évolutivité et la confiance.
- Combiner agents personnalisés et solutions prêtes à l'emploi. Pour exploiter pleinement le potentiel transformateur des agents d'IA, les organisations doivent aller au-delà de la simple activation d'agents intégrés à des suites logicielles. Ces agents prêts à l'emploi peuvent certes rationaliser les flux de travail routiniers, mais ils génèrent rarement un avantage stratégique. Pour tirer pleinement parti de l'IA, il est nécessaire de développer des agents sur mesure pour les processus à fort impact, tels que la résolution de bout en bout des problèmes clients, l'orchestration adaptative de la chaîne d'approvisionnement ou la prise de décision complexe. Ces agents doivent être parfaitement alignés sur la logique, les flux de données et les leviers de création de valeur de l'entreprise, ce qui les rend difficiles à reproduire et particulièrement performants.
- Rester agile face à l'évolution rapide des technologies. L'IA agentique est un domaine technologique émergent, et les solutions évoluent à un rythme soutenu. Les agents devront prendre en charge les flux de travail sur plusieurs systèmes et ne devront pas être intégrés de manière rigide à une plateforme spécifique. Une architecture évolutive et indépendante des fournisseurs est donc indispensable.
- Ces défis ne peuvent être relevés par la simple superposition de nouveaux composants, tels que des systèmes de stockage de mémoire ou des moteurs d'orchestration, sur les architectures d'IA existantes. Si ces capacités sont nécessaires, elles ne suffisent pas. Un changement architectural fondamental s'impose : passer d'une infrastructure statique, centrée sur les modèles logiques, à un environnement dynamique, modulaire et gouverné,

conçu spécifiquement pour l'intelligence multi-agents – le maillage d'IA multi-agents.

-
- Le maillage d'IA agentique est un paradigme architectural composable, distribué et indépendant des fournisseurs, qui permet à de multiples agents de raisonner, de collaborer et d'agir de manière autonome sur un large éventail de systèmes, d'outils et de modèles de langage — de façon sécurisée, à grande échelle et conçu pour évoluer avec la technologie. Ce paradigme repose sur cinq principes de conception qui se renforcent mutuellement :
-
- Composabilité. Tout agent, outil ou LLM peut être intégré au maillage sans modification du système.
- Intelligence distribuée. Les tâches peuvent être décomposées et résolues par des réseaux d'agents coopérants.
- Découplage par couches. Les fonctions logiques, de mémoire, d'orchestration et d'interface sont découplées afin de maximiser la modularité.
- Neutralité vis-à-vis des fournisseurs. Tous les composants peuvent être mis à jour ou remplacés indépendamment au gré des évolutions technologiques, évitant ainsi la dépendance à un fournisseur unique et pérennisant l'architecture. En particulier, les standards ouverts tels que le Model Context Protocol (MCP) et Agent2Agent (A2A) sont privilégiés par rapport aux protocoles propriétaires.
- Autonomie encadrée. Le comportement des agents est contrôlé de manière proactive via des politiques, des autorisations et des mécanismes d'escalade intégrés qui garantissent un fonctionnement sûr et transparent.
-
- Partager
-
- barre latérale
- Sept capacités interconnectées du maillage d'agents d'IA
- Le maillage d'IA agentique sert de couche de connexion et d'orchestration, permettant aux écosystèmes d'agents intelligents à grande échelle de fonctionner de manière sûre et efficace, et d'évoluer en continu. Il permet aux entreprises de coordonner des agents sur mesure et des agents prêts à l'emploi au sein d'un cadre unifié, de faciliter la collaboration multi-agents en leur permettant de partager le contexte et de déléguer des tâches, et d'atténuer les principaux risques tels que la prolifération des agents, la dérive d'autonomie et le manque d'observabilité, tout en préservant l'agilité nécessaire à une évolution technologique rapide (voir l'encadré « Sept capacités interconnectées du maillage d'IA agentique »).
-

-
- Partager
-
- barre latérale
- Modèles fondamentaux pour les agents : cinq exigences
- Au-delà de cette évolution architecturale, les organisations devront également revoir leurs stratégies LLM. Chaque agent personnalisé repose sur un modèle fondamental : le moteur de raisonnement qui sous-tend la perception, la prise de décision et l'interaction. À l'ère des agents, les exigences imposées aux LLM évoluent considérablement. Les agents ne sont plus de simples copilotes passifs ; ce sont des systèmes autonomes, persistants et embarqués. Il en résulte cinq catégories critiques d'exigences LLM, chacune associée à des contextes de déploiement spécifiques, pour lesquels différents types de modèles seront pertinents (voir l'encadré « Modèles fondamentaux pour agents : cinq exigences »).
-
- Enfin, pour déployer véritablement des agents à l'échelle de l'entreprise, les systèmes d'entreprise eux-mêmes doivent également évoluer.
-
- À court terme, les API (protocoles permettant à différentes applications logicielles de communiquer et d'échanger des données) resteront l'interface principale d'interaction entre les agents et les systèmes d'entreprise. Mais à long terme, les API seules ne suffiront plus. Les organisations doivent repenser leurs architectures informatiques autour d'un modèle centré sur l'agent, où les interfaces utilisateur, la logique et les couches d'accès aux données sont conçues nativement pour l'interaction machine plutôt que pour la navigation humaine. Dans un tel modèle, les systèmes ne sont plus organisés autour d'écrans et de formulaires, mais autour d'interfaces lisibles par machine, de flux de travail autonomes et de processus de décision pilotés par les agents.
-
- Cette transformation est déjà en cours. Microsoft intègre des agents au cœur de Dynamics 365 et de Microsoft 365 via Copilot Studio ; Salesforce étend Agentforce en une couche d'orchestration multi-agents ; SAP repense sa plateforme technologique d'entreprise (BTP) pour prendre en charge l'intégration d'agents via Joule. Ces changements annoncent une transition plus large : l'avenir des logiciels d'entreprise ne repose pas uniquement sur l'IA, mais sur une approche nativement axée sur les agents.
-
- Le principal défi ne sera pas technique, il sera humain.
- À mesure que les agents évoluent de copilotes passifs à acteurs proactifs – et se déploient à l'échelle de l'entreprise – la complexité qu'ils introduisent

sera non seulement technique, mais surtout organisationnelle. Le véritable défi réside dans la coordination, le jugement et la confiance. Cette complexité organisationnelle se manifestera principalement selon trois axes : la manière dont humains et agents cohabitent au quotidien ; la façon dont les organisations mettent en place une gouvernance pour les systèmes autonomes ; et la manière dont elles préviennent la prolifération incontrôlée des agents à mesure que leur création se démocratise.

-
- Cohabitation humain-agent. Les agents ne se contenteront pas d'assister les humains ; ils agiront à leurs côtés. Ceci soulève des questions complexes sur l'interaction et la coexistence : quand un agent doit-il prendre l'initiative ? Quand doit-il s'en remettre à l'humain ? Comment préserver l'autonomie et le contrôle humains sans compromettre les avantages mêmes qu'apportent les agents ? Clarifier ces rôles nécessitera du temps, des expérimentations et une adaptation culturelle. La confiance ne reposera pas uniquement sur les performances techniques ; elle dépendra de la transparence de la communication des agents, de la prévisibilité de leur comportement et de leur intégration intuitive aux flux de travail quotidiens.
- Contrôle de l'autonomie. Ce qui fait la force des agents – leur capacité à agir indépendamment – introduit également de l'ambiguïté. Contrairement aux outils traditionnels, les agents n'attendent pas d'instructions. Ils réagissent, s'adaptent et parfois surprennent. Apprivoiser cette nouvelle réalité implique de gérer les cas limites : que se passe-t-il si un agent agit de manière trop agressive ? Ou s'il omet de signaler un problème subtil ? L'enjeu n'est pas de supprimer l'autonomie, mais de la rendre intelligible et conforme aux attentes de l'organisation. Cette conformité ne sera pas figée. Elle devra évoluer au fur et à mesure que les agents apprennent, que les systèmes changent et que la confiance se renforce. Les mécanismes de contrôle doivent également prendre en compte le risque d'hallucinations, c'est-à-dire de résultats plausibles mais inexacts que les agents peuvent produire.

-
- Saisir l'avantage de l'IA agentive
- 13 juin 2025| Rapport
-
-
- Share
- Print
-
- Download
- Sauvegarder

- Un guide pratique pour les PDG afin de résoudre le paradoxe de l'IA générationnelle et de débloquer un impact à grande échelle grâce aux agents d'IA.
- TÉLÉCHARGEMENTS
- Rapport complet (28 pages)
- En un coup d'œil
- Chapitre 1
- Chapitre 2
-
- Chapitre 3
- Conclusion
- En un coup d'œil
- Près de huit entreprises sur dix déclarent utiliser l'IA de nouvelle génération, mais tout autant ne constatent aucun impact significatif sur leurs résultats financiers.1Considérez cela comme le « paradoxe de l'IA générationnelle ».
-
- Signature
- À propos des auteurs
- Au cœur de ce paradoxe se trouve un déséquilibre entre les copilotes et les chatbots « horizontaux » (à l'échelle de l'entreprise) — qui se sont développés rapidement mais n'offrent que des gains diffus et difficiles à mesurer — et les cas d'utilisation « verticaux » (spécifiques à une fonction) plus transformateurs — dont environ 90 % restent bloqués en mode pilote.
- Les agents d'IA offrent une solution au paradoxe de l'IA généraliste. En effet, ils ont le potentiel d'automatiser des processus métier complexes — en combinant autonomie, planification, mémoire et intégration — transformant ainsi l'IA généraliste d'un outil réactif en un collaborateur virtuel proactif et orienté vers un objectif.
- Ce changement permet bien plus qu'une simple amélioration de l'efficacité. Les agents décuplent l'agilité opérationnelle et créent de nouvelles sources de revenus.
- Mais pour exploiter pleinement le potentiel de l'IA agentielle, il ne suffit pas d'intégrer des agents aux flux de travail existants. Il faut repenser ces flux de travail de A à Z, en plaçant les agents au cœur du système.
-
- Partager
-
- barre latérale
- Avant-propos

- Un nouveau paradigme d'architecture d'IA – le maillage d'IA agentique – est nécessaire pour encadrer l'évolution rapide du paysage de l'IA organisationnelle et permettre aux équipes de combiner agents sur mesure et agents prêts à l'emploi, tout en maîtrisant la dette technique croissante et les nouveaux types de risques. Mais le plus grand défi ne sera pas technique. Il sera humain : gagner la confiance, favoriser l'adoption et mettre en place une gouvernance adéquate pour gérer l'autonomie des agents et prévenir leur prolifération incontrôlée.
- Pour amplifier leur impact à l'ère de l'intelligence artificielle, les organisations doivent repenser leurs approches de transformation par l'IA : passer d'initiatives dispersées à des programmes stratégiques ; des cas d'utilisation aux processus métier ; des équipes d'IA cloisonnées aux équipes de transformation transversales ; et de l'expérimentation à une mise en œuvre industrialisée et évolutive.
- Les organisations devront également mettre en place les fondements nécessaires pour opérer efficacement à l'ère des agents. Elles devront développer les compétences de leurs employés, adapter leur infrastructure technologique, accélérer la valorisation des données et déployer des mécanismes de gouvernance spécifiques aux agents. Le moment est venu de clore le chapitre de l'expérimentation en IA générale — un tournant que seul le PDG peut opérer.
- Chapitre 1
- Le paradoxe de l'IA générationnelle : déploiement généralisé, impact minimal
- Points clés
- Passez à la section suivante
-
-
- Partager
- Près de huit entreprises sur dix ont déployé l'IA de nouvelle génération sous une forme ou une autre, mais à peu près le même pourcentage déclare qu'elle n'a eu aucun impact significatif sur ses bénéfices.1 Nous appelons cela le « paradoxe de l’IA générale ».
- Le principal problème réside dans le déséquilibre entre les cas d'usage « horizontaux » et « verticaux ». Les premiers, comme les copilotes employés et les chatbots, sont largement déployés mais leurs bénéfices restent diffus, tandis que les cas d'usage verticaux, ou spécifiques à une fonction, à plus fort impact, dépassent rarement le stade pilote en raison d'obstacles techniques, organisationnels, liés aux données et culturels.
- À moins que les entreprises ne s'attaquent à ces obstacles, la promesse transformationnelle de l'IA de nouvelle génération restera largement inexploitée.

- L'intelligence artificielle de nouvelle génération est partout, sauf dans les comptes de résultat des entreprises.
-
- Partager
-
- barre latérale
- À propos de QuantumBlack, l'IA de McKinsey
- Avant même l'avènement de l'IA de nouvelle génération, l'intelligence artificielle s'était déjà taillé une place essentielle dans les entreprises, en alimentant des capacités avancées de prédiction, de classification et d'optimisation. Son potentiel de valeur était déjà estimé à un montant immense, entre 11 et 18 billions de dollars à l'échelle mondiale.2— principalement dans les domaines du marketing (permettant des fonctionnalités telles que le ciblage personnalisé des e-mails et la segmentation client), des ventes (qualification des prospects) et de la chaîne d'approvisionnement (optimisation des stocks et prévision de la demande). Pourtant, l'IA restait largement l'apanage des experts. De ce fait, son adoption par l'ensemble des employés était généralement lente. De 2018 à 2022, par exemple, l'adoption de l'IA est restée relativement stable, environ 50 % des entreprises déployant cette technologie dans une seule fonction métier, selon une étude de McKinsey (Graphique 1).
-
- Pièce justificative 1
- L'intelligence artificielle de génération a globalement accéléré le déploiement de l'IA.
- Nous nous efforçons de garantir l'accessibilité de notre site web aux personnes en situation de handicap. Pour toute information complémentaire concernant ce contenu, n'hésitez pas à nous contacter par courriel à l'adresse suivante : McKinsey_Website_Accessibility@mckinsey.com
- L'IA de nouvelle génération a étendu la portée de l'IA traditionnelle dans trois domaines révolutionnaires : la synthèse d'informations, la génération de contenu et la communication en langage humain. McKinsey estime que cette technologie pourrait générer entre 2 600 et 4 400 milliards de dollars de valeur ajoutée, en plus de celle déjà présente dans l'IA analytique traditionnelle.3
-
- Deux ans et demi après le lancement de ChatGPT, l'IA nouvelle génération a profondément transformé la manière dont les entreprises interagissent avec l'IA. Son potentiel de transformation réside non seulement dans les nouvelles fonctionnalités qu'elle introduit, mais aussi dans sa capacité à démocratiser l'accès aux technologies d'IA avancées au sein des organisations. Cette démocratisation a entraîné une forte augmentation de

la sensibilisation à l'IA et des expérimentations associées : selon la dernière enquête mondiale de McKinsey sur l'IA ,4Plus de 78 % des entreprises utilisent désormais l'IA de nouvelle génération dans au moins une fonction commerciale (contre 55 % un an auparavant).

-
- Toutefois, cet enthousiasme ne s'est pas encore traduit par des résultats économiques concrets. Plus de 80 % des entreprises déclarent toujours que leurs initiatives en matière d'IA de nouvelle génération n'ont aucun impact significatif sur leurs bénéfices.5De plus, seulement 1 % des entreprises que nous avons interrogées considèrent leurs stratégies en matière d'IA de génération comme matures .6On pourrait parler du « paradoxe de l'IA générationnelle » : malgré toute l'énergie, les investissements et le potentiel entourant cette technologie, son impact à grande échelle ne s'est pas encore concrétisé pour la plupart des organisations.
-
- Au cœur du paradoxe de l'IA générationnelle se trouve un déséquilibre entre les cas d'utilisation horizontaux et verticaux.
- De nombreuses organisations ont déployé des cas d'utilisation horizontaux, tels que des copilotes et des chatbots à l'échelle de l'entreprise ; près de 70 % des entreprises du classement Fortune 500, par exemple, utilisent Microsoft 365 Copilot.7Ces outils sont généralement perçus comme des leviers d'amélioration de la productivité individuelle, permettant aux employés de gagner du temps sur les tâches routinières et d'accéder à l'information et de la synthétiser plus efficacement. Toutefois, ces améliorations, bien que réelles, ont tendance à être peu visibles parmi les employés. Par conséquent, elles ne se traduisent pas facilement par des résultats concrets, que ce soit au niveau du chiffre d'affaires ou des bénéfices.
-
- À l'inverse, les cas d'usage verticaux — ceux intégrés à des fonctions et processus métier spécifiques — ont connu un déploiement limité dans la plupart des entreprises, malgré leur potentiel plus élevé d'impact économique direct (voir graphique 2). Selon une étude de McKinsey, moins de 10 % des cas d'usage déployés dépassent le stade du projet pilote .8Même une fois pleinement déployées, ces solutions n'ont généralement pris en charge que des étapes isolées d'un processus métier et ont fonctionné de manière réactive, sur intervention humaine, plutôt que de façon proactive ou autonome. De ce fait, leur impact sur la performance de l'entreprise est resté limité.
-
- Pièce n° 2

- Dans tous les domaines fonctionnels de l'entreprise, les cas d'utilisation de l'IA de nouvelle génération se répartissent généralement en deux catégories : horizontale et verticale.
- Nous nous efforçons de garantir l'accessibilité de notre site web aux personnes en situation de handicap. Pour toute information complémentaire concernant ce contenu, n'hésitez pas à nous contacter par courriel à l'adresse suivante : McKinsey_Website_Accessibility@mckinsey.com
- Comment expliquer ce déséquilibre ? D’une part, les solutions de copilotage déployées horizontalement, telles que Microsoft Copilot ou Google AI Workspace, sont accessibles et prêtes à l’emploi, et relativement faciles à mettre en œuvre. (Dans de nombreux cas, activer Microsoft Copilot se résume à ajouter une extension à un contrat Office 365 existant, sans nécessiter de refonte des flux de travail ni d’efforts majeurs de gestion du changement.) Le déploiement rapide des chatbots d’entreprise a également été motivé par des préoccupations liées à la réduction des risques. Face à l’expérimentation par les employés de modèles de langage externes de grande taille (LLM) comme ChatGPT, de nombreuses organisations ont mis en place des alternatives internes et sécurisées afin de limiter les fuites de données et de garantir la conformité aux politiques de sécurité de l’entreprise.
-
- Gros plan sur des brins de cordes colorées s'entremêlant.
- Repenser l'entreprise grâce à la technologie et à l'IA : une transformation radicale.
- Mardi 21 avril, de 11 h à 11 h 30 HAE / de 17 h à 17 h 30 HEC
-
- Rejoignez Robert Levin et Kate Smaje , associés principaux de McKinsey et co-auteurs de « Rewired : The McKinsey Playbook on How Leading Companies Win with Technology and AI », pour une conversation sur ce que font les entreprises leaders pour transformer leurs activités grâce à l'IA.
-
-
- Inscrivez-vous ici
- Le déploiement limité et la portée restreinte des cas d'utilisation verticaux peuvent être attribués à six facteurs principaux :
-
- Initiatives fragmentées. Dans de nombreuses entreprises, les cas d'usage verticaux ont été identifiés par une approche ascendante et très granulaire au sein de chaque fonction. De fait, moins de 30 % des entreprises indiquent que leur PDG pilote directement leur stratégie en matière d'IA .9Cela a entraîné une prolifération de micro-initiatives déconnectées et une

dispersion des investissements en IA, avec une coordination limitée au niveau de l'entreprise.

- Absence de solutions éprouvées et prêtes à l'emploi. Contrairement aux applications horizontales standard, telles que les copilotes, les cas d'usage verticaux nécessitent souvent un développement sur mesure. De ce fait, les équipes sont fréquemment contraintes de tout construire de A à Z, en utilisant des technologies émergentes et en constante évolution avec lesquelles elles ont une expérience limitée. Si de nombreuses entreprises ont investi dans des data scientists pour développer des modèles d'IA, elles manquent souvent d'ingénieurs MLOps, pourtant essentiels pour industrialiser, déployer et maintenir ces modèles en production.
- Limitations technologiques des LLM. Malgré leurs capacités impressionnantes, la première génération de LLM présentait des limitations qui ont considérablement freiné leur déploiement à grande échelle. Premièrement, les LLM peuvent produire des résultats inexacts, ce qui les rend peu fiables dans les environnements où la précision et la reproductibilité sont essentielles. De plus, malgré leur puissance, les LLM sont fondamentalement passifs ; ils n’agissent que sur commande et ne peuvent pas gérer les flux de travail de manière autonome ni prendre de décisions sans intervention humaine. Les LLM ont également éprouvé des difficultés à gérer des flux de travail complexes comportant plusieurs étapes, points de décision ou logiques de branchement. Enfin, de nombreux LLM actuels disposent d’une mémoire persistante limitée, ce qui rend difficile le suivi du contexte dans le temps ou un fonctionnement cohérent lors d’interactions prolongées.
- Équipes d'IA cloisonnées. Les centres d'excellence en IA ont joué un rôle crucial dans l'accélération de la sensibilisation et de l'expérimentation au sein de nombreuses organisations. Cependant, dans bien des cas, ces équipes ont fonctionné en silos, développant des modèles d'IA indépendamment des fonctions informatiques, de données ou métiers essentielles. Cette autonomie, bien qu'utile pour le prototypage rapide, a souvent rendu les solutions difficiles à déployer à grande échelle en raison d'une mauvaise intégration aux systèmes d'entreprise, de flux de données fragmentés ou d'un manque d'alignement opérationnel.
- Lacunes en matière d'accessibilité et de qualité des données. Ces lacunes concernent aussi bien les données structurées que non structurées, ces dernières restant largement non réglementées dans la plupart des organisations.
- Craintes culturelles et inertie organisationnelle. Dans de nombreuses organisations, le déploiement de l'IA s'est heurté à une résistance implicite de la part des équipes opérationnelles et du management intermédiaire, en raison de la crainte de perturbations, de l'incertitude quant à l'impact sur l'emploi et du manque de familiarité avec cette technologie.

- Malgré son impact limité sur les résultats financiers jusqu'à présent, la première vague d'IA de génération s'est avérée loin d'être vaine. Elle a enrichi les compétences des employés, permis une expérimentation à grande échelle, accéléré la familiarisation avec l'IA dans tous les services et aidé les organisations à développer des compétences essentielles en ingénierie rapide, en évaluation des modèles et en gouvernance. Tout cela a jeté les bases d'une seconde phase plus intégrée et transformatrice : l'avènement des agents d'IA .10
-
- Chapitre 2
- Du paradoxe à la récompense : comment les agents peuvent déployer l’IA à grande échelle
- Points clés
- Passez à la section suivante
-
-
- Partager
- En automatisant les processus métier complexes, les agents permettent d'exploiter pleinement le potentiel des cas d'usage verticaux. Les entreprises visionnaires tirent déjà parti de la puissance des agents pour transformer leurs processus clés.
- Pour exploiter pleinement le potentiel des agents, les entreprises doivent réinventer leur façon de travailler : modifier les flux de tâches, redéfinir les rôles humains et construire dès le départ des processus centrés sur les agents.
- Pour y parvenir, il faudra un nouveau paradigme d'architecture d'IA : le maillage d'IA agentif, capable d'intégrer des agents développés sur mesure et des agents prêts à l'emploi. Mais le plus grand défi ne sera pas technique. Il sera humain : gagner la confiance pour favoriser l'adoption et établir les protocoles de gouvernance appropriés.
- La percée : L'automatisation des flux de travail complexes permet d'exploiter pleinement le potentiel des cas d'utilisation verticaux.
- Les langages de modélisation linguistique (LLM) ont révolutionné la manière dont les organisations interagissent avec les données, permettant la synthèse d'informations, la génération de contenu et l'interaction en langage naturel. Cependant, malgré leur puissance, les LLM sont restés fondamentalement réactifs et isolés des systèmes d'entreprise, incapables de conserver en mémoire les interactions passées ou le contexte entre les sessions ou les requêtes. Leur rôle s'est longtemps limité à l'amélioration de la productivité individuelle par le biais de tâches isolées. Les agents d'IA marquent une évolution majeure de l'IA d'entreprise, faisant passer l'IA générative de la génération réactive de contenu à une exécution autonome et orientée vers un objectif. Ces agents peuvent comprendre les

objectifs, les décomposer en sous-tâches, interagir avec les humains et les systèmes, exécuter des actions et s'adapter en temps réel, le tout avec une intervention humaine minimale. Ils y parviennent en combinant les LLM avec des composants technologiques supplémentaires offrant des capacités de mémoire, de planification, d'orchestration et d'intégration.

-
- Grâce à ces nouvelles fonctionnalités, les agents d'IA étendent le potentiel des solutions horizontales, transformant les copilotes généralistes d'outils passifs en collaborateurs proactifs. Ces derniers ne se contentent plus de répondre aux sollicitations, mais surveillent également les tableaux de bord, déclenchent des flux de travail, assurent le suivi des actions en cours et fournissent des informations pertinentes en temps réel. Toutefois, la véritable avancée réside dans le domaine vertical, où l'IA agentique permet l'automatisation de flux de travail métier complexes impliquant de multiples étapes, acteurs et systèmes – des processus qui dépassaient auparavant les capacités des outils d'IA de première génération.
-
- Les agents apportent bien plus que de l'efficacité : ils décuplent l'agilité opérationnelle et ouvrent de nouvelles perspectives de revenus.
- Côté opérations, les agents prennent en charge les tâches routinières et gourmandes en données, permettant ainsi aux humains de se concentrer sur des missions à plus forte valeur ajoutée. Mais leur rôle ne s'arrête pas là : ils transforment les processus de cinq manières :
-
- Les agents accélèrent l'exécution en éliminant les délais entre les tâches et en permettant le traitement parallèle. Contrairement aux flux de travail traditionnels qui reposent sur des transferts séquentiels, les agents peuvent coordonner et exécuter plusieurs étapes simultanément, réduisant ainsi le temps de cycle et améliorant la réactivité.
- Les agents apportent l'adaptabilité. En ingérant des données en continu, ils peuvent ajuster les flux de processus en temps réel, en réorganisant les séquences de tâches, en réattribuant les priorités ou en signalant les anomalies avant qu'elles ne provoquent des défaillances. Les flux de travail sont ainsi non seulement plus rapides, mais aussi plus intelligents.
- Les agents permettent une personnalisation. En adaptant les interactions et les décisions aux profils ou comportements individuels des clients, les agents peuvent ajuster le processus de manière dynamique afin de maximiser la satisfaction et les résultats.
- Les agents apportent de la flexibilité aux opérations. Étant numériques, leur capacité d'exécution peut augmenter ou diminuer en temps réel en fonction de la charge de travail, de la saisonnalité de l'activité ou des pics d'activité imprévus – une flexibilité difficile à obtenir avec des modèles de ressources humaines fixes.

- Les agents contribuent également à renforcer la résilience des opérations. En surveillant les perturbations, en réacheminant les opérations et en intervenant uniquement en cas de besoin, ils assurent la continuité des processus, qu'il s'agisse de chaînes d'approvisionnement confrontées à des retards portuaires ou de flux de travail de service s'adaptant aux pannes de système.
- Dans un environnement de chaîne d'approvisionnement complexe, un agent d'IA pourrait, par exemple, jouer le rôle d'une couche d'orchestration autonome pour les opérations d'approvisionnement, d'entreposage et de distribution. Connecté aux systèmes internes (tels que le système de planification de la chaîne d'approvisionnement ou le système de gestion d'entrepôt) et aux sources de données externes (comme les prévisions météorologiques, les flux de données des fournisseurs et les signaux de la demande), cet agent pourrait prévoir la demande en continu. Il pourrait ensuite identifier les risques, tels que les retards ou les perturbations, et replanifier dynamiquement les flux de transport et de stocks. En sélectionnant le mode de transport optimal en fonction du coût, du délai et de l'impact environnemental, l'agent pourrait réallouer les stocks entre les entrepôts, négocier directement avec les systèmes externes et remonter les décisions nécessitant une expertise stratégique. Résultat : une amélioration des niveaux de service, une réduction des coûts logistiques et une diminution des émissions.
-
- Les agents peuvent également contribuer à stimuler la croissance du chiffre d'affaires en amplifiant les sources de revenus existantes et en en débloquant de nouvelles :
-
- Augmenter les revenus existants. Dans le e-commerce, des agents intégrés aux boutiques en ligne ou aux applications pourraient analyser proactivement le comportement des utilisateurs, le contenu de leur panier et le contexte (par exemple, la saisonnalité ou l'historique d'achats) afin de proposer des offres de vente incitative et de vente croisée en temps réel. Dans le secteur financier, ces agents pourraient aider les clients à trouver des produits financiers adaptés à leurs besoins, tels que des prêts, des assurances ou des portefeuilles d'investissement, en leur fournissant des conseils personnalisés en fonction de leur profil financier, de leur situation personnelle et de leurs habitudes de consommation.
- Création de nouvelles sources de revenus. Pour les entreprises industrielles, des agents intégrés aux produits ou équipements connectés pourraient surveiller l'utilisation, détecter les seuils de performance et débloquer automatiquement des fonctionnalités ou déclencher des interventions de maintenance, permettant ainsi des modèles de rémunération à l'usage, par abonnement ou basés sur la performance. De

même, les entreprises de services pourraient intégrer leur expertise interne (raisonnement juridique, interprétation fiscale et bonnes pratiques d'approvisionnement) dans des agents d'IA proposés sous forme de logiciels en tant que service (SaaS) ou d'API à leurs clients, partenaires ou PME ne disposant pas de cette expertise en interne.

- En résumé, l'IA agentielle ne se contente pas d'automatiser. Elle redéfinit la manière dont les organisations fonctionnent, s'adaptent et créent de la valeur.
-
- Ce n'est plus de la science-fiction : les entreprises visionnaires exploitent le pouvoir des agents
- Les études de cas suivantes démontrent comment QuantumBlack aide les organisations à constituer des effectifs d'agents, avec des résultats qui vont bien au-delà des gains d'efficacité.
-
- Étude de cas 1 : Comment une banque a utilisé des « usines numériques » hybrides pour moderniser ses applications existantes
- Le problème : une grande banque devait moderniser son système central informatique, composé de 400 logiciels – un projet colossal dont le budget dépassait les 600 millions de dollars. De grandes équipes de développeurs ont abordé le projet en effectuant des tâches manuelles et répétitives, ce qui a engendré des difficultés de coordination entre les différents services. Elles s'appuyaient également sur une documentation et un code souvent lents et sujets aux erreurs. Si les outils d'IA de première génération ont permis d'accélérer certaines tâches, la progression est restée lente et laborieuse.
-
- L'approche par agents : les employés humains ont été promus à des rôles de supervision, encadrant des équipes d'agents d'IA. Chaque équipe contribue à un objectif commun selon une séquence définie (voir l'illustration 3). Ces équipes documentent a posteriori l'application existante, écrivent du nouveau code, examinent le code des autres agents et intègrent ce code dans des fonctionnalités qui sont ensuite testées par d'autres agents avant la livraison du produit final. Libérés des tâches manuelles répétitives, les superviseurs humains guident chaque étape du processus, améliorant ainsi la qualité des livrables et réduisant le nombre de sprints nécessaires à l'implémentation de nouvelles fonctionnalités.
-
- Impact : Réduction de plus de 50 % du temps et des efforts consacrés par les équipes pionnières.
-
- Pièce justificative 3

- Une grande banque a modernisé son infrastructure technologique traditionnelle avec une usine numérique hybride IA-humain.
- Nous nous efforçons de garantir l'accessibilité de notre site web aux personnes en situation de handicap. Pour toute information complémentaire concernant ce contenu, n'hésitez pas à nous contacter par courriel à l'adresse suivante : McKinsey_Website_Accessibility@mckinsey.com
- Étude de cas 2 : Comment un cabinet d'études a amélioré la qualité de ses données pour obtenir des informations plus approfondies sur le marché
- Le problème : une société d'études de marché et de veille stratégique consacrait des ressources considérables à garantir la qualité des données, s'appuyant sur une équipe de plus de 500 personnes chargées de collecter, structurer et codifier les données, puis de générer des analyses personnalisées pour ses clients. Ce processus, réalisé manuellement, était sujet aux erreurs, dont 80 % étaient détectées par les clients eux-mêmes.
-
- L'approche multi-agents : une solution autonome identifie les anomalies de données et explique les variations des ventes ou des parts de marché. Elle analyse les signaux internes, tels que les modifications de la nomenclature des produits, et les événements externes identifiés par des recherches web, comme les rappels de produits ou les intempéries. Les facteurs les plus influents sont synthétisés, hiérarchisés et mis à la disposition des décideurs. Grâce à la recherche avancée et au raisonnement contextuel, les agents font souvent émerger des informations qu'il serait difficile pour des analystes humains de découvrir manuellement. Bien que le système ne soit pas encore en production, il est pleinement fonctionnel et a démontré un fort potentiel pour libérer les analystes de leurs tâches stratégiques.
-
- Impact : Gain de productivité potentiel de plus de 60 % et économies attendues de plus de 3 millions de dollars par an.
-
- Étude de cas 3 : Comment une banque a repensé sa méthode de création de notes d'évaluation des risques de crédit
- Le problème : les chargés de clientèle d'une banque de détail passaient des semaines à rédiger et à peaufiner des notes d'évaluation du risque de crédit afin de faciliter leurs décisions et de se conformer aux exigences réglementaires (voir l'annexe 4). Ce processus les obligeait à examiner et à extraire manuellement des informations provenant d'au moins dix sources de données différentes, et à élaborer un raisonnement complexe et nuancé sur des sections interdépendantes, par exemple l'évolution conjointe des prêts, des revenus et de la trésorerie.
-

- L’approche par agents : En étroite collaboration avec les experts en risque de crédit et les chargés de clientèle de la banque, une preuve de concept a été développée afin de transformer le flux de travail des notes de crédit grâce à des agents d’IA. Ces agents assistent les chargés de clientèle en extrayant des données, en rédigeant des sections de notes, en générant des scores de confiance pour prioriser les analyses et en suggérant des questions de suivi pertinentes. Dans ce modèle, le rôle de l’analyste évolue de la rédaction manuelle vers la supervision stratégique et la gestion des exceptions.
-
- Impact : Une augmentation potentielle de la productivité de 20 à 60 %, dont une amélioration de 30 % du délai de traitement des crédits.
-
- Pièce justificative 4
- Une banque de détail a utilisé des agents d'IA pour réinventer le processus de création de notes d'information sur le risque de crédit.
- Nous nous efforçons de garantir l'accessibilité de notre site web aux personnes en situation de handicap. Pour toute information complémentaire concernant ce contenu, n'hésitez pas à nous contacter par courriel à l'adresse suivante : McKinsey_Website_Accessibility@mckinsey.com
- Pour maximiser la valeur des agents d'IA, il faut réinventer les processus.
- Pour exploiter pleinement le potentiel de l'IA dans les secteurs verticaux, il ne suffit pas d'intégrer des agents aux flux de travail existants. Il est nécessaire de repenser entièrement la conception, en passant de l'automatisation des tâches au sein d'un processus existant à la réinvention du processus dans son intégralité, avec la collaboration d'humains et d'agents. En effet, lorsque des agents sont intégrés à un processus existant sans refonte, ils servent généralement d'assistants plus rapides : ils génèrent du contenu, extraient des données ou exécutent des étapes prédéfinies. Mais le processus lui-même reste séquentiel, rigide et soumis aux contraintes humaines.
-
- Repenser un processus autour d'agents ne se limite pas à automatiser les flux de travail existants ; cela implique de repenser entièrement l'architecture du flux de tâches. Il s'agit notamment de réorganiser les étapes, de redistribuer les responsabilités entre humains et agents, et de concevoir le processus pour exploiter pleinement les atouts de l'IA agentielle : exécution parallèle réduisant considérablement les délais, adaptabilité en temps réel aux conditions changeantes, personnalisation poussée à grande échelle et capacité flexible s'ajustant instantanément à la demande.
-

- Prenons l'exemple d'un centre d'appels clients hypothétique. Avant l'introduction d'agents IA, ce centre utilisait des outils d'IA générale pour assister le personnel d'assistance humain : recherche d'articles dans les bases de connaissances, synthèse de l'historique des tickets et aide à la rédaction des réponses. Si cette assistance a permis d'accélérer le processus et de réduire la charge cognitive, le processus lui-même restait entièrement manuel et réactif, les agents humains gérant toujours chaque étape du diagnostic, de la coordination et de la résolution. Le potentiel d'amélioration de la productivité était modeste, avec généralement une augmentation du temps de résolution et de la productivité de 5 à 10 %.
-
- Imaginez maintenant qu'un centre d'appels intègre des agents IA tout en conservant en grande partie son flux de travail actuel : des agents sont ajoutés pour intervenir à des étapes spécifiques sans que l'acheminement, le suivi ou la résolution des demandes ne soient modifiés de bout en bout. Ces agents peuvent classer les tickets, suggérer les causes probables, proposer des solutions et même résoudre de manière autonome les problèmes fréquents et simples (comme la réinitialisation des mots de passe). Si l'impact peut être accru – avec un gain de temps estimé entre 20 et 40 % et une réduction du volume de demandes en attente de 30 à 50 % – , les difficultés de coordination et la faible adaptabilité empêchent des gains véritablement significatifs.
-
- Mais le véritable changement s'opère au troisième niveau, lorsque le processus du centre d'appels est repensé autour de l'autonomie des agents. Dans ce modèle, les agents IA ne se contentent pas de répondre : ils détectent proactivement les problèmes clients courants (tels que les retards de livraison, les échecs de paiement ou les interruptions de service) en analysant les tendances sur tous les canaux, anticipent les besoins probables, initient automatiquement les solutions (comme les remboursements, les commandes de nouveaux articles ou la mise à jour des informations de compte) et communiquent directement avec les clients par chat ou e-mail. Les agents humains sont repositionnés comme gestionnaires d'escalade et responsables de la qualité du service, et interviennent uniquement lorsque les agents détectent une incertitude ou une exception aux schémas habituels. L'impact à ce niveau est transformateur. Cela pourrait permettre une amélioration radicale de la productivité du service client. Jusqu'à 80 % des incidents courants pourraient être résolus de manière autonome, avec une réduction du temps de résolution de 60 à 90 % (voir l'illustration 5).
-
- Pièce à conviction 5

- Les agents détiennent la clé de la percée, à condition que les processus soient réinventés et non simplement optimisés.
- Nous nous efforçons de garantir l'accessibilité de notre site web aux personnes en situation de handicap. Pour toute information complémentaire concernant ce contenu, n'hésitez pas à nous contacter par courriel à l'adresse suivante : McKinsey_Website_Accessibility@mckinsey.com
- Bien sûr, tous les processus métier ne nécessitent pas une refonte complète. Une simple automatisation des tâches suffit pour les flux de travail standardisés et répétitifs, à faible variabilité – comme le traitement de la paie, l'approbation des notes de frais ou la réinitialisation des mots de passe – où les gains proviennent principalement de la réduction des interventions manuelles. En revanche, les processus complexes, transversaux, sujets aux exceptions ou étroitement liés à la performance de l'entreprise justifient souvent une refonte totale. Parmi les indicateurs clés d'une telle refonte figurent une lourdeur de coordination, des séquences rigides qui ralentissent la réactivité, des interventions humaines fréquentes pour des décisions qui pourraient être basées sur les données, et des possibilités d'adaptation dynamique ou de personnalisation. Dans ces cas, repenser le processus en s'appuyant sur la capacité de l'agent à orchestrer, s'adapter et apprendre apporte une valeur ajoutée bien supérieure à une simple accélération des flux de travail existants.
-
- Un nouveau paradigme d'architecture d'IA — le maillage d'IA agentique — est nécessaire pour orchestrer la création de valeur à l'ère de l'agentivité.
- Pour déployer leurs agents à grande échelle, les entreprises devront relever un triple défi : gérer les nouveaux risques liés aux agents IA, combiner des systèmes d’agents personnalisés et prêts à l’emploi, et rester agiles face à l’évolution rapide des technologies (tout en évitant la dépendance vis-à-vis d’un fournisseur unique).
-
- Gérer une nouvelle vague de risques. Les agents introduisent une nouvelle catégorie de risques systémiques que les architectures d'IA traditionnelles, conçues principalement pour des cas d'usage isolés et centrés sur la gestion des actifs numériques, n'ont jamais été conçues pour gérer : autonomie incontrôlée, accès fragmenté au système, manque d'observabilité et de traçabilité, surface d'attaque croissante, prolifération et duplication des agents. Ce qui commence comme une automatisation intelligente peut rapidement se transformer en chaos opérationnel, à moins de reposer sur des fondements privilégiant le contrôle, l'évolutivité et la confiance.
- Combiner agents personnalisés et solutions prêtes à l'emploi. Pour exploiter pleinement le potentiel transformateur des agents d'IA, les

organisations doivent aller au-delà de la simple activation d'agents intégrés à des suites logicielles. Ces agents prêts à l'emploi peuvent certes rationaliser les flux de travail routiniers, mais ils génèrent rarement un avantage stratégique. Pour tirer pleinement parti de l'IA, il est nécessaire de développer des agents sur mesure pour les processus à fort impact, tels que la résolution de bout en bout des problèmes clients, l'orchestration adaptative de la chaîne d'approvisionnement ou la prise de décision complexe. Ces agents doivent être parfaitement alignés sur la logique, les flux de données et les leviers de création de valeur de l'entreprise, ce qui les rend difficiles à reproduire et particulièrement performants.

- Rester agile face à l'évolution rapide des technologies. L'IA agentique est un domaine technologique émergent, et les solutions évoluent à un rythme soutenu. Les agents devront prendre en charge les flux de travail sur plusieurs systèmes et ne devront pas être intégrés de manière rigide à une plateforme spécifique. Une architecture évolutive et indépendante des fournisseurs est donc indispensable.
- Ces défis ne peuvent être relevés par la simple superposition de nouveaux composants, tels que des systèmes de stockage de mémoire ou des moteurs d'orchestration, sur les architectures d'IA existantes. Si ces capacités sont nécessaires, elles ne suffisent pas. Un changement architectural fondamental s'impose : passer d'une infrastructure statique, centrée sur les modèles logiques, à un environnement dynamique, modulaire et gouverné, conçu spécifiquement pour l'intelligence multi-agents – le maillage d'IA multi-agents.
-
- Le maillage d'IA agentique est un paradigme architectural composable, distribué et indépendant des fournisseurs, qui permet à de multiples agents de raisonner, de collaborer et d'agir de manière autonome sur un large éventail de systèmes, d'outils et de modèles de langage — de façon sécurisée, à grande échelle et conçu pour évoluer avec la technologie. Ce paradigme repose sur cinq principes de conception qui se renforcent mutuellement :
-
- Composabilité. Tout agent, outil ou LLM peut être intégré au maillage sans modification du système.
- Intelligence distribuée. Les tâches peuvent être décomposées et résolues par des réseaux d'agents coopérants.
- Découplage par couches. Les fonctions logiques, de mémoire, d'orchestration et d'interface sont découplées afin de maximiser la modularité.
- Neutralité vis-à-vis des fournisseurs. Tous les composants peuvent être mis à jour ou remplacés indépendamment au gré des évolutions technologiques, évitant ainsi la dépendance à un fournisseur unique et

pérennisant l'architecture. En particulier, les standards ouverts tels que le Model Context Protocol (MCP) et Agent2Agent (A2A) sont privilégiés par rapport aux protocoles propriétaires.

- Autonomie encadrée. Le comportement des agents est contrôlé de manière proactive via des politiques, des autorisations et des mécanismes d'escalade intégrés qui garantissent un fonctionnement sûr et transparent.
-
- Partager
-
- barre latérale
- Sept capacités interconnectées du maillage d'agents d'IA
- Le maillage d'IA agentique sert de couche de connexion et d'orchestration, permettant aux écosystèmes d'agents intelligents à grande échelle de fonctionner de manière sûre et efficace, et d'évoluer en continu. Il permet aux entreprises de coordonner des agents sur mesure et des agents prêts à l'emploi au sein d'un cadre unifié, de faciliter la collaboration multi-agents en leur permettant de partager le contexte et de déléguer des tâches, et d'atténuer les principaux risques tels que la prolifération des agents, la dérive d'autonomie et le manque d'observabilité, tout en préservant l'agilité nécessaire à une évolution technologique rapide (voir l'encadré « Sept capacités interconnectées du maillage d'IA agentique »).
-
-
- Partager
-
- barre latérale
- Modèles fondamentaux pour les agents : cinq exigences
- Au-delà de cette évolution architecturale, les organisations devront également revoir leurs stratégies LLM. Chaque agent personnalisé repose sur un modèle fondamental : le moteur de raisonnement qui sous-tend la perception, la prise de décision et l'interaction. À l'ère des agents, les exigences imposées aux LLM évoluent considérablement. Les agents ne sont plus de simples copilotes passifs ; ce sont des systèmes autonomes, persistants et embarqués. Il en résulte cinq catégories critiques d'exigences LLM, chacune associée à des contextes de déploiement spécifiques, pour lesquels différents types de modèles seront pertinents (voir l'encadré « Modèles fondamentaux pour agents : cinq exigences »).
-
- Enfin, pour déployer véritablement des agents à l'échelle de l'entreprise, les systèmes d'entreprise eux-mêmes doivent également évoluer.
-

- À court terme, les API (protocoles permettant à différentes applications logicielles de communiquer et d'échanger des données) resteront l'interface principale d'interaction entre les agents et les systèmes d'entreprise. Mais à long terme, les API seules ne suffiront plus. Les organisations doivent repenser leurs architectures informatiques autour d'un modèle centré sur l'agent, où les interfaces utilisateur, la logique et les couches d'accès aux données sont conçues nativement pour l'interaction machine plutôt que pour la navigation humaine. Dans un tel modèle, les systèmes ne sont plus organisés autour d'écrans et de formulaires, mais autour d'interfaces lisibles par machine, de flux de travail autonomes et de processus de décision pilotés par les agents.
-
- Cette transformation est déjà en cours. Microsoft intègre des agents au cœur de Dynamics 365 et de Microsoft 365 via Copilot Studio ; Salesforce étend Agentforce en une couche d'orchestration multi-agents ; SAP repense sa plateforme technologique d'entreprise (BTP) pour prendre en charge l'intégration d'agents via Joule. Ces changements annoncent une transition plus large : l'avenir des logiciels d'entreprise ne repose pas uniquement sur l'IA, mais sur une approche nativement axée sur les agents.

Le principal défi ne sera pas technique, il sera humain.

- À mesure que les agents évoluent de copilotes passifs à acteurs proactifs – et se déploient à l'échelle de l'entreprise – la complexité qu'ils introduisent sera non seulement technique, mais surtout organisationnelle. Le véritable défi réside dans la coordination, le jugement et la confiance. Cette complexité organisationnelle se manifestera principalement selon trois axes : la manière dont humains et agents cohabitent au quotidien ; la façon dont les organisations mettent en place une gouvernance pour les systèmes autonomes ; et la manière dont elles préviennent la prolifération incontrôlée des agents à mesure que leur création se démocratise.
- **Cohabitation humain-agent**. Les agents ne se contenteront pas d'assister les humains ; ils agiront à leurs côtés. Ceci soulève des questions complexes sur l'interaction et la coexistence : quand un agent doit-il prendre l'initiative ? Quand doit-il s'en remettre à l'humain ? Comment préserver l'autonomie et le contrôle humains sans compromettre les avantages mêmes qu'apportent les agents ? Clarifier ces rôles nécessitera du temps, des expérimentations et une adaptation culturelle. La confiance ne reposera pas uniquement sur les performances techniques ; elle dépendra de la transparence de la communication des agents, de la

prévisibilité de leur comportement et de leur intégration intuitive aux flux de travail quotidiens.

- **Contrôle de l'autonomie**. Ce qui fait la force des agents – leur capacité à agir indépendamment – introduit également de l'ambiguïté. Contrairement aux outils traditionnels, les agents n'attendent pas d'instructions. Ils réagissent, s'adaptent et parfois surprennent. Apprivoiser cette nouvelle réalité implique de gérer les cas limites : que se passe-t-il si un agent agit de manière trop agressive ? Ou s'il omet de signaler un problème subtil ? L'enjeu n'est pas de supprimer l'autonomie, mais de la rendre intelligible et conforme aux attentes de l'organisation. Cette conformité ne sera pas figée. Elle devra évoluer au fur et à mesure que les agents apprennent, que les systèmes changent et que la confiance se renforce. Les mécanismes de contrôle doivent également prendre en compte le risque d'hallucinations, c'est-à-dire de résultats plausibles mais inexacts que les agents peuvent produire.
- **Maîtriser la prolifération des agents**. Comme aux débuts de l'automatisation robotisée des processus (RPA), le risque de prolifération des agents est bien réel : une multiplication incontrôlée d'agents redondants, fragmentés et non gouvernés au sein des équipes et des fonctions. Avec les plateformes low-code et no-code qui rendent la création d'agents accessible à tous, les organisations s'exposent à une nouvelle forme d'informatique parallèle : des agents qui se multiplient entre les équipes, dupliquent les efforts ou fonctionnent sans supervision. Comment éviter cette fragmentation ? Qui décide des agents à développer et de ceux à supprimer ? Sans gouvernance structurée, sans normes de conception et sans gestion du cycle de vie, les écosystèmes d'agents peuvent rapidement devenir fragiles, redondants et non évolutifs.

Les agents permettent de libérer tout le potentiel des cas d'usage verticaux, offrant aux entreprises la possibilité de générer de la valeur bien au-delà des simples gains d'efficacité. Mais pour concrétiser ce potentiel, il est nécessaire de repenser la transformation par l'IA : une approche adaptée à la nature unique des agents et capable de pallier les limitations persistantes qu'ils ne peuvent résoudre à eux seuls. Cette approche fera l'objet du chapitre suivant.

Chapitre 3

La transformation par l'IA à un tournant décisif : le rôle du PDG à l'ère de l'agentivité

Points clés

- Pour générer un impact à l'ère de l'agentivité, les organisations doivent passer d'initiatives dispersées à des programmes stratégiques ; de cas

d'utilisation à des processus métier ; d'équipes d'IA cloisonnées à des équipes de transformation transversales ; et de l'expérimentation à une mise en œuvre industrialisée et évolutive.

- Pour déployer davantage d'agents, les organisations devront également établir de nouvelles bases en renforçant les compétences de leur personnel, en adaptant leur infrastructure technologique et en développant de nouvelles structures de gouvernance pour les agents
- Le moment est venu de mettre fin à la phase d'expérimentation de l'IA de nouvelle génération – un tournant que seul le PDG peut opérer.

Pour amplifier son impact à l'ère de l'agentivité, il est nécessaire de repenser l'approche de transformation de l'IA.

Contrairement aux outils d'IA classiques qui s'intègrent facilement aux flux de travail existants, les agents d'IA exigent une transformation plus profonde, impliquant une refonte des processus métier et une intégration poussée aux systèmes d'entreprise. McKinsey propose une stratégie éprouvée, Rewired, pour les transformations pilotées par l'IA .11Pour tirer pleinement parti de cette opportunité d'action, les organisations doivent s'appuyer sur cette base et remodeler fondamentalement leur approche de transformation par l'IA selon quatre dimensions :

- **Stratégie** : Des initiatives tactiques éparses aux programmes stratégiques. Face à l'IA agentielle qui s'apprête à redéfinir les fondements de la concurrence, les organisations doivent dépasser l'identification des cas d'usage de base et aligner directement leurs initiatives d'IA sur leurs priorités stratégiques les plus critiques. Cela implique non seulement de traduire les objectifs existants – tels que l'amélioration de l'efficacité opérationnelle, le renforcement de la relation client ou le développement de la conformité – en domaines de transformation accessibles à l'IA, mais aussi d'adopter une vision prospective. Les dirigeants doivent inciter leurs équipes à repenser le modèle opérationnel actuel et à explorer comment l'IA peut être utilisée pour réinventer des pans entiers de l'activité, créer de nouvelles sources de revenus et bâtir des avantages concurrentiels qui définiront le leadership au cours de la prochaine décennie.
- **Unité de transformation :** du cas d’usage aux processus métier. Lors de la première vague d’adoption de l’IA, la plupart des initiatives sectorielles se concentraient sur l’intégration d’une solution à une étape spécifique

d’un processus existant, ce qui se traduisait généralement par des gains limités sans modifier la structure globale du travail. Avec les agents IA, le paradigme change radicalement. L’opportunité réside désormais non plus dans l’optimisation de tâches isolées, mais dans la transformation de processus métier entiers grâce à l’intégration d’agents tout au long de la chaîne de valeur. Par conséquent, les initiatives d’IA ne doivent plus se limiter à un seul cas d’usage, mais viser la refonte complète d’un processus ou d’un parcours utilisateur. Dans les secteurs d’activité, cela signifie passer de la question : « Où puis-je utiliser l’IA dans cette fonction ? » à : « À quoi ressemblerait cette fonction si des agents en géraient 60 % ? » Cela implique de repenser les flux de travail, la logique de décision, les interactions homme-machine et les indicateurs de performance

- **Modèle de déploiement** : Des équipes IA cloisonnées aux équipes de transformation transversales. Les centres d’excellence en IA ont joué un rôle clé dans l’accélération de la sensibilisation et de l’expérimentation de l’IA au sein des organisations. Cependant, ce modèle atteint ses limites à l’ère des agents, où ces derniers sont profondément intégrés aux systèmes d’entreprise, opèrent sur des processus métier complexes et dépendent de données de haute qualité comme principal carburant. Dans ce contexte, les initiatives d’IA ne peuvent plus être menées par des équipes IA isolées et spécialisées. Pour réussir à grande échelle, les organisations doivent adopter un modèle de déploiement transversal, s’appuyant sur des équipes de transformation pérennes composées d’experts métier, de concepteurs de processus, d’ingénieurs IA et MLOps, d’architectes informatiques, d’ingénieurs logiciels et d’ingénieurs de données.
- Processus de mise en œuvre : De l’expérimentation à un déploiement industrialisé et évolutif. Si la phase précédente s’est légitimement concentrée sur l’exploration du potentiel de l’IA de nouvelle génération, les organisations doivent désormais adopter un modèle de déploiement industrialisé, où les solutions sont conçues dès le départ pour évoluer, tant sur le plan technique que financier. Cela implique d’anticiper l’ensemble des prérequis techniques pour un déploiement en entreprise, notamment en termes d’intégration système, de surveillance quotidienne et de gestion des mises en production, mais aussi d’estimer rigoureusement les coûts d’exploitation futurs et de concevoir une solution permettant de les minimiser. Contrairement aux systèmes informatiques traditionnels, dont

les coûts d’exploitation annuels représentent généralement 10 à 20 % des coûts de développement initiaux, ce modèle exige des coûts d’exploitation réduits. Les solutions d'IA de nouvelle génération, notamment à grande échelle, peuvent engendrer des coûts récurrents supérieurs à l'investissement initial. La conception pour l'évolutivité doit donc prendre en compte non seulement la robustesse technique, mais aussi la viabilité économique, en particulier pour les applications à fort volume.

Quatre facteurs clés de succès sont nécessaires pour opérer efficacement à l'ère de l'agentivité.

Repenser l'approche de la transformation par l'IA est une étape importante, mais insuffisante. Pour exploiter pleinement leur potentiel à grande échelle, les organisations doivent également activer un ensemble solide de leviers favorisant les changements structurels, culturels et techniques nécessaires à l'intégration des agents dans les opérations quotidiennes. Ces leviers s'articulent autour de quatre dimensions : les personnes, la gouvernance, l'architecture technologique et les données. Chacune de ces dimensions constitue un socle pour un déploiement évolutif, sécurisé et performant des agents à l'échelle de l'entreprise.

- **Ressources humaines** : Former les équipes et créer de nouveaux rôles. Les équipes doivent être préparées aux nouvelles méthodes de travail axées sur la collaboration humain-agent. Cela implique de promouvoir une approche « humain + agent » par le biais d’un changement culturel, de formations ciblées et du soutien aux pionniers en tant qu’ambassadeurs internes. De nouveaux rôles doivent également être créés, tels que des ingénieurs de réponse rapide pour optimiser les interactions, des orchestrateurs d’agents pour gérer les flux de travail des agents et des concepteurs intégrant l’humain dans le processus pour gérer les exceptions et instaurer la confiance.
- **Gouvernance** : Garantir le contrôle de l’autonomie et prévenir la prolifération des agents. L’essor des agents autonomes engendre la nécessité d’une gouvernance robuste afin d’éviter les risques et une prolifération incontrôlée. Les entreprises doivent définir des cadres de gouvernance établissant les niveaux d’autonomie des agents, les limites de

décision, la surveillance des comportements et les mécanismes d'audit. Les politiques de développement, de déploiement et d'utilisation doivent également être formalisées, de même que les systèmes de classification regroupant les agents par fonction (par exemple, les automatisateurs de tâches, les orchestrateurs de domaine et les collaborateurs virtuels), chacun étant doté d'un modèle de supervision approprié.

- **Architecture technologique** : Établir les fondements de l'interopérabilité et de la mise à l'échelle. Les agents, qu'ils soient développés sur mesure ou prêts à l'emploi, doivent fonctionner au sein d'un écosystème fragmenté de systèmes, de données et de flux de travail. À court terme, les organisations doivent faire évoluer leur architecture d'IA, passant de configurations centrées sur les modèles de logique métier (LLM) à un maillage d'IA basé sur les agents. Au-delà de cette première étape, elles doivent anticiper l'architecture de nouvelle génération, dans laquelle tous les systèmes d'entreprise seront repensés autour des agents, tant au niveau de l'interface utilisateur que de la logique métier et des opérations quotidiennes.
- **Données** : Accélérer la commercialisation des données et combler les lacunes en matière de qualité des données non structurées. Enfin, les acteurs dépendent de la qualité et de l'accessibilité des données d'entreprise. Les organisations doivent passer de pipelines de données spécifiques à chaque cas d'usage à des produits de données réutilisables et étendre la gouvernance des données aux données non structurées.

Les PDG sont confrontés à un défi de leadership : mettre fin à la phase d'expérimentation de l'IA de nouvelle génération

L'essor des agents d'IA représente bien plus qu'une simple évolution technologique. Ces agents constituent un tournant stratégique majeur qui redéfinira le fonctionnement, la compétitivité et la création de valeur des entreprises. Pour réussir cette transition, les organisations doivent dépasser le stade de l'expérimentation et des projets pilotes et s'engager dans une nouvelle phase de transformation à grande échelle, à l'échelle de l'entreprise.

Ce virage stratégique ne peut être délégué ; il doit être initié et piloté par le PDG. Il reposera sur trois actions clés :

- **Action 1** : Conclure la phase d'expérimentation et réaligner les priorités en matière d'IA. Procéder à une analyse structurée pour tirer les enseignements de l'expérience, abandonner les projets pilotes non reproductibles à grande échelle et clore officiellement la phase exploratoire. Recentrer les efforts sur les programmes d'IA stratégiques ciblant les domaines et processus à fort impact.
- **Action 2** : Repenser la gouvernance et le modèle opérationnel de l'IA. Mettre en place un conseil stratégique de l'IA réunissant les dirigeants, le directeur des ressources humaines, le directeur des données et le directeur des systèmes d'information. Ce conseil devra superviser la stratégie en matière d'IA, coordonner les investissements dans l'IA, les technologies de l'information et les données, et mettre en œuvre des mécanismes rigoureux de suivi de la valeur, basés sur des indicateurs clés de performance liés aux résultats de l'entreprise.
- **Action 3** : Lancer un premier projet pilote de transformation et initialiser simultanément l'infrastructure technologique de l'IA agentique. Initier un nombre restreint de transformations de flux de travail à fort impact, pilotées par l'IA agentique, dans les domaines d'activité clés. En parallèle, préparer le terrain pour une infrastructure technologique d'IA agentique en investissant dans les leviers essentiels : infrastructure technologique, qualité des données, cadres de gouvernance et préparation des équipes.

Action 1 : Conclure la phase d'expérimentation et réaligner les priorités en matière d'IA. Procéder à une analyse structurée pour tirer les enseignements de l'expérience, abandonner les projets pilotes non reproductibles à grande échelle et clore officiellement la phase exploratoire. Recentrer les efforts sur les programmes d'IA stratégiques ciblant les domaines et processus à fort impact.

Action 2 : Repenser la gouvernance et le modèle opérationnel de l'IA. Mettre en place un conseil stratégique de l'IA réunissant les dirigeants, le directeur des ressources humaines, le directeur des données et le directeur des systèmes d'information. Ce conseil devra superviser la stratégie en matière d'IA, coordonner les investissements dans l'IA, les technologies de l'information et les données, et mettre en œuvre des mécanismes rigoureux de suivi de la valeur, basés sur des indicateurs clés de performance liés aux résultats de l'entreprise.

Action 3 : Lancer un premier projet pilote de transformation et initialiser simultanément l’infrastructure technologique de l’IA agentique. Initier un nombre restreint de transformations de flux de travail à fort impact, pilotées par l’IA agentique, dans les domaines d’activité clés. En parallèle, préparer le terrain pour une infrastructure technologique d’IA agentique en investissant dans les leviers essentiels : infrastructure technologique, qualité des données, cadres de gouvernance et préparation des équipes.

Conclusion

Comme toute technologie véritablement disruptive, les agents d'IA ont le pouvoir de bouleverser la donne. Bien conçus, ils offrent aux entreprises en retard une occasion unique de rattraper leur retard et de redéfinir leur compétitivité. Mal conçus – ou pas du tout – ils risquent d'accélérer le déclin des leaders actuels du marché. Nous sommes à un tournant stratégique.

Bien que la technologie continue d'évoluer, elle est déjà suffisamment mature pour impulser des changements profonds et transformateurs dans tous les secteurs. Cependant, pour concrétiser pleinement le potentiel de l'IA agentielle, les dirigeants doivent repenser leur approche de la transformation par l'IA : non pas par une série de projets pilotes épars, mais par des efforts de refonte globale et ciblée. Cela implique d'identifier les quelques domaines d'activité à fort potentiel et d'exploiter tous les leviers : de la réinvention des flux de travail à la redistribution des tâches entre humains et machines, en passant par la restructuration de l'organisation selon de nouveaux modèles opérationnels.

Certains dirigeants sont déjà passés à l'action, non seulement en déployant des flottes d'agents, mais aussi en restructurant leurs organisations pour exploiter pleinement leur potentiel de transformation. (Moderna, par exemple, a fusionné ses directions des ressources humaines et des technologies de l'information.)13—ce qui indique que l'IA n'est pas qu'un simple outil technique, mais une force transformatrice pour le monde du travail. Il s'agit d'une évolution structurelle vers une nouvelle forme d'entreprise. L'IA agentique n'est pas une simple évolution, elle constitue le fondement du modèle opérationnel de demain. Les PDG qui agissent dès maintenant ne se contenteront pas d'améliorer leurs performances. Ils redéfiniront la manière dont leurs organisations pensent, décident et agissent.

L'heure de l'exploration touche à sa fin. L'heure de la transformation a sonné.