KPMG

Maîtriser les risques de l'IA agentique

Novembre 2025

KPMG. Make The Difference.*
*KPMG. Pour Faire la Différence.

I. Avant-Propos

Nous sommes déjà rentrés dans une nouvelle ère, celle de l'IA agentique. Les agents IA ne se contentent plus de produire du contenu comme les IA génératives : ils agissent. Grâce à leurs capacités de raisonnement, de planification, de mémoire et d'interaction avec des outils, ils exécutent des tâches complexes avec un niveau d'autonomie élevé. Ce changement de paradigme – de la production de contenu à l'action sur le monde réel et virtuel – redéfinit la nature même des opportunités, mais aussi des risques pour toute organisation.

L'agentification de processus promet des gains majeurs sur l'ensemble de la chaîne de valeur. Qu'ils opèrent comme spécialistes de tâches ciblées, collaborateurs augmentant les équipes ou orchestrateurs de systèmes complexes, les agents déploient leur potentiel dans toutes les fonctions et tous les secteurs. Les bénéfices attendus sont multiples : gains d'efficacité et de productivité, agilité renforcée, résilience accrue et, in fine, une expérience profondément améliorée pour les clients comme pour les collaborateurs.

Cependant les caractéristiques des agents IA introduisent des risques d'une nature nouvelle et amplifient des risques existants, bien plus difficiles à anticiper et maîtriser que ceux associés aux logiciels traditionnels. La surface d'attaque devient plus large et dynamique. Les organisations qui développent et déploient des agents IA doivent maitriser des multiples risques, par exemple des dysfonctionnements causés par des hallucinations, des attaques menées par des acteurs malveillants sur la mémoire et les outils, ou des comportements émergents d'agents IA, non alignés sur les valeurs humaines.

Face à ces enjeux, les organisations doivent anticiper la transformation qui permettra une intégration optimale des agents IA dans les processus, sur la base de trois piliers : une gouvernance solide, des dispositifs techniques robustes et une supervision, humaine ou automatisée, efficace.

C'est pour préparer cette transition et construire une IA agentique de confiance que KPMG détaille dans cette étude les différentes catégories de risques liées aux agents IA, identifie les moyens de réduction des risques envisageables et propose une feuille de route articulée en chantiers concrets pour chaque grande fonction de l'entreprise (cybersécurité, risques, RH, etc.).

Vincent Maret
Associé, KPMG Advisory
Risk Consulting - Cybersécurité et AI Trust
vmaret@kpmg.fr

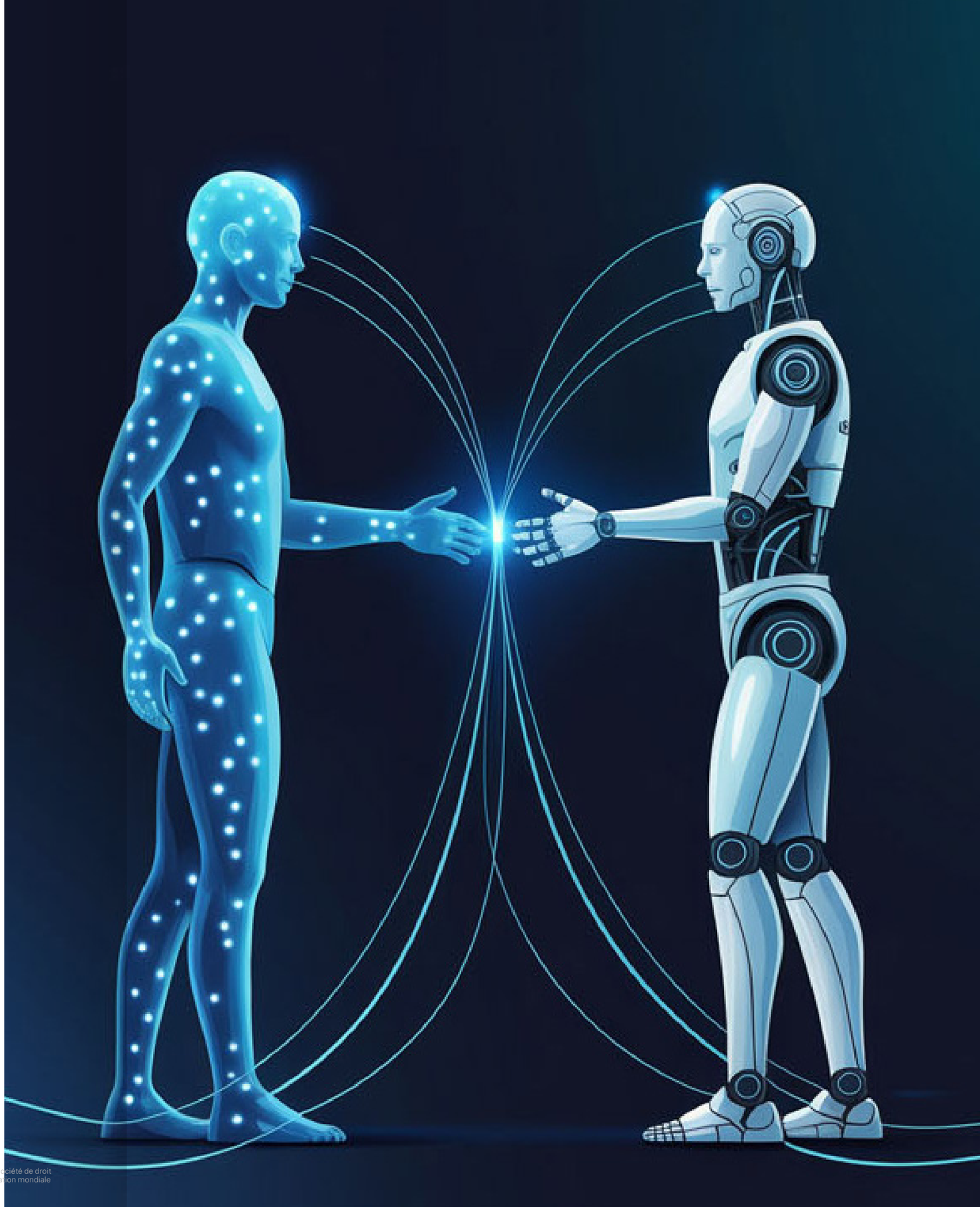


Table des matières

II. Qu'est-ce qu'un agent IA ?

Le 30 novembre 2022, le grand public découvrait les capacités impressionnantes de ChatGPT en matière de dialogue, de réponse à des questions, de génération de contenus, de résumé et de traduction de texte. Cette avancée majeure était le fruit d'une décennie d'innovations technologiques, notamment la représentation vectorielle des mots, l'architecture « Transformer », et les premiers LLM (Large Langage Models) que furent BERT et GPT-1. Loin de s'essouffler, cette vague d'innovation s'est poursuivie avec une succession rapide de nouveaux modèles encore plus performants d'OpenAI (GPT-4o, o3, o4), Anthropic (Claude), Google (Gemini), Meta (Llama), Mistral (Le Chat), xAi (Grok) ou Deepseek (R1).

Là où depuis des décennies, l'IA traditionnelle se focalisait sur la production de résultats prédictifs (scores, classifications ou recommandations), l'IA générative s'est imposée très rapidement en produisant du contenu riche (texte, image, vidéo, code). Les plus puissants de ces modèles, dits « de fondation », disposent de capacités couvrant une très large diversité de sujets et de cas d'usages. Ils sont entrainés à partir d'immenses masses de données, à des coûts colossaux, ce qui crée une forte incitation à réutiliser ces modèles sur étagère.

Aujourd'hui les LLM évoluent vers une nouvelle catégorie d'IA, les agents IA. S'il n'existe pas de définition officielle et universellement reconnue de ce qu'est un « agent IA », on peut décrire leurs caractéristiques principales.

1. Réflexion et planification

Les agents IA sont des systèmes d'IA qui s'appuient sur les capacités de réflexion et de raisonnement des LLM sous-jacents pour atteindre un objectif. Ils fonctionnent selon quatre étapes clefs :

Planification

L'agent élabore un plan permettant d'atteindre le but défini par l'utilisateur, en s'appuyant sur le contexte (mémoire, perception de l'environnement) ainsi que sur son rôle et/ou sa personnalité. Ce plan est décomposé en tâches logiques nécessaires à la réalisation de l'objectif.

Critique

L'agent analyse son propre plan, challenge sa capacité à atteindre l'objectif, et l'adapte le cas échéant.

Éxécution

L'agent exécute les tâches en utilisant des outils ou en collaborant avec d'autres agents spécialisés.

Vérification

L'agent analyse les résultats pour s'assurer que l'objectif est bien atteint.

2. Mémoire

Contrairement aux LLM classiques, dont la mémoire se limite à l'historique d'une seule conversation, les agents IA possèdent une mémoire à long terme. Une telle capacité leur permet de conserver des contextes, des préférences, des expériences, des feedbacks ou des données de référence liés à l'utilisateur, à l'environnement ou aux tâches précédemment réalisées. Grâce à cette mémoire, un agent peut notamment :

- **Optimiser** son comportement en fonction des informations reçues de son environnement suite à des actions ;
- **Adapter** son comportement aux préférences ou aux habitudes de chaque utilisateur ;
- **Mener** à bien des tâches complexes s'étalant sur plusieurs étapes.

3. Outils

Là où un simple LLM génère du contenu en fonction d'un prompt et de la connaissance accumulée dans ses paramètres, un agent IA peut utiliser des outils pour atteindre ses objectifs. L'agent choisit l'outil adéquat (appel à une API, accès à un site web ou à une base de données, exécution de code, etc.), lui envoie une requête, puis intègre le résultat obtenu dans le prompt soumis au LLM sous-jacent pour générer la réponse finale. Des tels outils peuvent être utilisés par les LLM pour :

- **Augmenter leurs connaissances** : interrogation de sites web, d'API ou de bases de données internes ou externes, de RAG ;
- **Étendre leurs capacités** : calculatrice, interpréteur de code, calendrier, génération d'images ou de schémas ;
- **Percevoir leur environnement et agir dessus** : interaction avec le système sur lequel l'agent est hébergé (notamment lecture de fichiers et exécution de commandes), collecte de données provenant de capteurs, lecture et mise à jour de bases de données, envoi d'e-mails, paiements, ordres envoyés à un robot, etc.).

Selon les cas, un agent peut également solliciter un autre agent, plus spécialisé, pour réaliser une tâche (voir ci-après.).

4. Autonomie

Contrairement à un LLM, principalement conçu pour générer du texte en réponse à un prompt, un agent IA s'appuie sur des frameworks logiciels spécialisés, comme LangChain, LlamaIndex, CrewAI, MetaGPT ou AutoGen, qui orchestrent le cycle itératif « Perception – Raisonnement – Action ». Ce cycle peut être déclenché par une demande explicite (instruction d'un utilisateur) ou par un évènement contextuel (réception d'une facture, alerte système, etc.).

Grâce à cette approche, un agent IA peut prendre des décisions, s'adapter et gérer des situations nouvelles ou complexes, avec une supervision humaine minimale.

5. IA agentique : révolution ou évolution ?

Les agents IA ne sont qu'une évolution des LLM. Les modèles les plus avancés d'OpenAI, Anthropic, Google, Meta, Mistral ou Deepseek disposent aujourd'hui pour la plupart de capacités de recours à des outils, de mémoire de long terme et de réflexion/planification. En outre, des extensions comme « Computer Use » d'Anthropic et « Operator » ou « ChatGPT agent » d'OpenAI offrent des capacités d'autonomie accrues, permettant notamment d'interagir avec des interfaces graphiques comme des navigateurs Web.

En revanche, l'ampleur des capacités des agents IA créent les conditions d'une révolution dans l'utilisation de l'IA dans les organisations.

III. Des modalités de mise en œuvre variées

1. Architectures

Il existe deux grands types d'architecture de systèmes d'IA agentique :

mono-agent

Un agent unique gère seul toutes les étapes (raisonnement, planification, utilisation d'outils et dialogue avec l'utilisateur) pour atteindre l'objectif fixé.

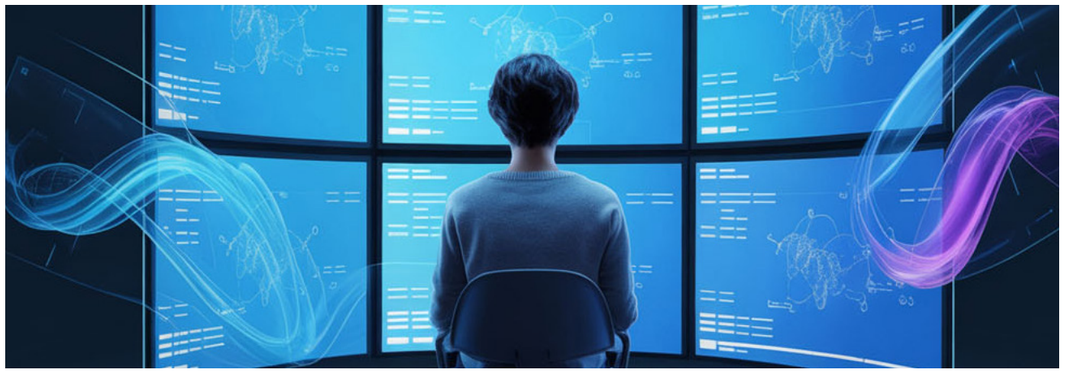
multi-agents

Plusieurs agents spécialisés avec des capacités spécifiques, des rôles voire des « personnalités » propres, agissent en collaboration ou en compétition dans le même environnement.

A. Architecture multi-agents

Plusieurs configurations sont envisageables dans un environnement multi-agents :

Séquentielle	Les agents interviennent les uns après les autres, selon un processus prédéfini. Chaque agent s'appuie sur le résultat du précédent et transmet le sien au suivant.
Hiérarchique (ou verticale)	Un agent « manager » planifie la stratégie permettant d'atteindre l'objectif, délègue les tâches opérationnelles à des agents « exécutants » spécialisés, supervise leurs actions et consolide leurs résultats. Cette structure peut comporter plusieurs niveaux de management.
Collaborative (ou horizontale)	Des agents coopèrent sans supervision centrale, en échangeant des informations pour atteindre un but commun. Les agents peuvent se voir attribuer des « personnalités » (ex : créatif, critique, prudent) pour enrichir le processus.
Compétitive	Plusieurs agents réalisent la même tâche en parallèle. Un superviseur (humain ou un autre agent) sélectionne ensuite le meilleur résultat ou en combine plusieurs.



B. Modes opératoires

Les agents IA peuvent opérer selon plusieurs modes, qu'ils fonctionnent de manière autonome ou qu'ils soient intégrés à des applications ou plateformes existantes. Le framework TACO (« Tasker, Automator, Collaborator, Orchestrator ») de KPMG considère 4 modes opératoires, avec des niveaux de complexité croissants :

Les "exécutants"

Ils exécutent des tâches uniques, bien définies et répétitives. Par exemple, vérifier qu'une facture contient toutes les informations obligatoires.

Les "automatiseurs"

En réponse à des événements spécifiques, ils déclenchent des actions de manière autonome ou semi-autonome pour atteindre un but prédéfini. Par exemple, générer des commandes pour réapprovisionner un stock.

Les "collaborateurs"

Ils assistent directement des utilisateurs. Sur la base d'une instruction ou d'un objectif donné, ils exécutent les tâches demandées et restituent le résultat.

Les "orchestrateurs"

Sur la base d'objectifs de haut niveau, ils coordonnent des opérations complexes en pilotant d'autres agents, des applications, des systèmes d'information et des humains.

2. Coopérations

La capacité de coopération des agents IA est une dimension importante de leur valeur ajoutée.

A. Coopération avec d'autres agents

Les agents IA peuvent collaborer entre eux dans différents contextes :

- Au sein d'une même organisation : par exemple, un agent responsable des paiements coopère avec un agent qui valide les factures ;
- Entre plusieurs organisations différentes : par exemple, l'agent d'un fournisseur négocie une vente avec l'agent d'un client ;
- Entre une organisation et un particulier : par exemple, l'assistant d'achat personnel d'un utilisateur interagit avec les agents d'un site e-commerce.

Des protocoles de communication dédiés comme A2A (Agent to Agent) permettent aux agents IA de dialoguer et de partager de manière structurée des informations essentielles : objectifs, plans d'action, résultats, connaissances ou encore feedbacks. Ces échanges peuvent aussi s'appuyer sur des mémoires partagées entre plusieurs agents.

B. Coopération avec les humains

Grâce aux LLM qui les animent, les agents excellent dans la communication avec les humains, qu'il s'agisse de collaborateurs ou de clients. Ils peuvent notamment :

- Prendre en compte des instruction détaillées et des objectifs complexes ;
- Demander des précisions sur les instructions ou les objectifs ;
- Adapter leur mode de communication à leurs utilisateurs et interlocuteurs ;
- Fournir des explications sur leurs actions et leurs décisions ;
- Prendre en compte les retours de l'utilisateur sur les actions réalisées et l'atteinte des objectifs.

C. Un nouvel écosystème hybride humains-agents IA

Avec l'IA agentique, les agents s'intègrent aux processus des organisations, interagissant avec les systèmes, les collaborateurs et d'autres agents. Dans ce nouveau cadre, le rôle de l'humain évolue : il devient un « manager d'agents » :

- En **définissant** les objectifs et les missions des agents ;
- En **sélectionnant** les bons agents pour chaque tâche ;
- En **donnant** les instructions et en fixant le niveau d'autonomie ;
- En **validant** les actions importantes et en fournissant de l'aide si besoin ;
- En **intervenant** en cas d'incident, de situation trop complexe ou d'exigences règlementaires ;
- En **supervisant** la performance globale des processus automatisés.

IV. Des cas d'usage multiples

Les agents IA sont susceptibles d'automatiser et d'optimiser de multiples processus métiers au sein des organisations. En voici quelques exemples :

1. Au sein des fonctions

Business Intelligence	• Identification de tendances sectorielles • Aide à la décision
Cybersécurité	• Détection des vulnérabilités et des intrusions • Réponse aux incidents
IT	• Support technique aux utilisateurs • Génération et optimisation de code
Juridique	• Analyse et rédaction de contrats • Recherche jurisprudentielle
Marketing	• Génération de contenu (textes, images) • Ciblage et personnalisation des messages
Productivité	• Planification de réunions et gestion d'agendas • Assistance administrative (notes de frais, etc.)
R & D	• Analyse de données expérimentales • Aide à la rédaction de publications scientifiques
Ressources Humaines	• Onboarding des nouveaux employés • Support et réponse aux questions des collaborateurs
Service Client	• Support client automatisé et gestion des réclamations • Analyse de la qualité des interactions
Supply Chain	• Optimisation des flux logistiques • Gestion des risques fournisseurs
Ventes	• Qualification et priorisation des prospects • Assistance à la négociation

2. Dans les secteurs

Agriculture	• Surveillance des cultures • Prédiction des rendements
Banque & Assurance	• Détection de fraude • Conseil en investissement personnalisé
Commerce de détail	• Personnalisation de l'expérience client • Optimisation des stocks et des réassorts
Construction	• Détection d'anomalies sur plans ou photos • Suivi de la conformité et des risques
Création & Médias	• Modération automatisée des plateformes • Assistance à la scénarisation
Éducation & Formation	• Tutorat et parcours d'apprentissage personnalisés • Aide à la correction et à l'évaluation
Énergie	• Optimisation de la consommation énergétique • Maintenance prédictive des infrastructures
Jeux & Divertissement	• Personnages non-joueurs (PNJ) au comportement réaliste • Personnalisation de l'expérience de jeu
Transport & Logistique	• Optimisation dynamique des itinéraires et des flottes • Gestion prédictive des retards et des incidents
Industrie	• Optimisation des processus de production • Maintenance prédictive
Santé	• Optimisation du parcours patient et des rendez-vous • Aide à la documentation clinique pour les praticiens
Télécoms	• Prédiction de l'attrition client (churn) • Support client intelligent et automatisé

4. Des promesses de gains significatifs pour les organisations

Performance et Efficacité	• Réduction des taux d'erreurs dans les tâches • Accélération des processus et des délais de traitement • Automatisation des tâches répétitives et à faible valeur ajoutée, réduisant les coûts opérationnels • Capacité d'optimisation de la performance en prenant en compte l'expérience accumulée
Continuité et résilience	• Fonctionnement 24 heures sur 24 • Capacité de traitement s'adaptant à la charge de travail (pics d'activité, saisonnalité) • En cas de défaillance, relai automatique par un autre agent • Capacité des agents à traiter des situations complexes ou peu courantes
Expérience Améliorée	• Pour les collaborateurs, temps libéré pour des missions plus stratégiques, créatives et innovantes • Pour les clients, service plus réactif, hyper-personnalisé et disponible à tout moment

V. Quels risques pour l'IA agentique ?

La capacité des agents IA à raisonner, prendre des décisions et agir de manière autonome constitue une évolution majeure qui introduit des risques importants. Leur comportement est plus difficile à anticiper, à tester et à encadrer que celui des logiciels traditionnels fondés sur des règles déterministes.

Avant de lancer un projet d'agent IA, il est donc crucial de mener une analyse de risques structurée afin d'examiner les vulnérabilités, d'identifier les scénarios de dysfonctionnement ou de malveillance et d'évaluer la gravité ainsi que la probabilité de chaque risque.

1. De nouveaux risques liés à l'IA agentique

L'IA générative et les LLM ont introduit des risques nouveaux, mais désormais bien connus, tels que les hallucinations, la génération de contenus toxiques ou l'injection de prompt. Cependant, l'avènement de l'IA agentique marque un tournant critique, amplifiant les risques existants et en créant de nouveaux. Ses capacités étendues, comme la planification ou l'utilisation d'outils, génèrent une surface d'attaque plus large, plus dynamique et bien plus difficile à maîtriser. On constate dans l'IA agentique des modes de défaillance, des comportements émergents et des vecteurs d'attaques qui n'existent ni dans l'univers des systèmes informatiques traditionnels et déterministes, ni dans celui de l'IA générative et des LLM.

En outre, l'IA agentique franchit un seuil fondamental : elle ne se limite plus à générer du texte, elle agit directement sur le monde réel ou virtuel. Ce changement de paradigme modifie profondément la nature du risque. Là où un LLM se limite à produire du contenu, un agent peut réaliser des actions concrètes, comme modifier une base de données ou effectuer un virement, avec potentiellement des impacts physiques, financiers ou organisationnels. Par ailleurs, au sein d'une organisation, un agent IA a accès à différents systèmes, applications, bases de données, API, etc., avec les privilèges nécessaires pour son rôle. Toute défaillance ou cyberattaque le visant peut donc gravement impacter le fonctionnement et la sécurité de l'entreprise, au-delà du process agentifié.

On peut distinguer quatre catégories de risques liés aux agents IA :

- Des défaillances et des dysfonctionnements accidentels des agents IA, sans aucune intention malveillante ;
- Des utilisateurs ou des attaquants qui manipulent les agents IA pour exécuter des actions nuisibles ou voler des données ;
- Des agents IA spécifiquement conçus pour agir de manière hostile ;
- Des agents IA qui agissent de façon contraire aux objectifs ou non alignée avec les valeurs humaines.

2. Quels risques liés aux capacités de raisonnement et de planification ?

Un agent IA s'appuie sur le LLM sous-jacent pour planifier ses actions. Or les LLM peuvent halluciner, commettre des erreurs de raisonnement ou mal utiliser leurs outils. Il existe donc un risque que le plan produit par l'agent pour atteindre l'objectif soit incorrect, ou que la décision prise par l'agent ne soit pas appropriée. On constate parfois que le raisonnement de l'agent « tourne en boucle », provoquant des actions répétitives absurdes, comme l'envoi de centaines d'emails. Ce risque est d'autant plus présent que les tâches données à l'agent IA et le contexte dans lequel il évolue sont complexes, ambiguës, évolutives.

Les capacités de raisonnement d'un agent IA peuvent par ailleurs être ciblées et détournées par un attaquant pour le forcer à prendre des décisions erronées ou à exécuter des actions malveillantes. Les principales méthodes d'attaque sont :

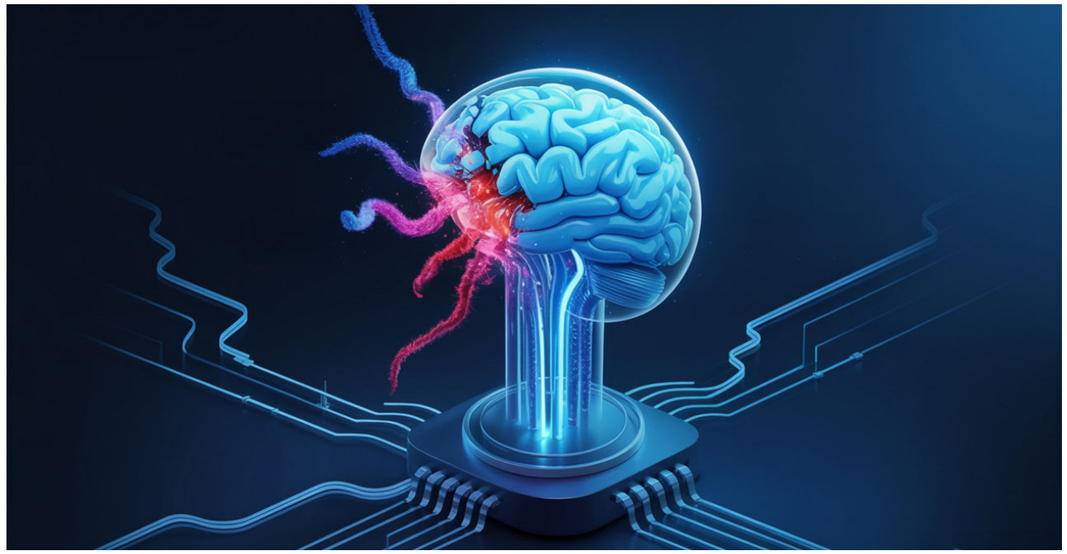
- **La manipulation des entrées** : l'attaquant utilise des techniques comme l'injection de prompt pour tromper l'agent.
- **L'empoisonnement du modèle** : une attaque plus sophistiquée consiste à corrompre les données d'apprentissage du LLM sous-jacent pour y insérer une « porte dérobée » (backdoor), permettant ensuite de manipuler secrètement le comportement de l'agent.

3. Quels risques liés aux capacités de mémoire ?

La mémoire d'un agent IA représente une surface d'attaque critique car son contenu peut influencer directement le comportement de l'agent. Les menaces sont de deux natures :

Corruption de la mémoire : un attaquant peut parvenir à injecter de « faux souvenirs » dans la mémoire de l'agent. Cette altération peut fausser le comportement de l'agent et l'amener à réaliser des actions inappropriées. Elle peut en outre persister pendant de longues périodes sans nouvelle intervention de l'attaquant.

Fuite de données : un attaquant peut utiliser des injections de prompt pour faire révéler par l'agent des informations sensibles contenues dans sa mémoire.



4. Quels risques liés à l'utilisation d'outils ?

La gestion des identités et des accès par un agent IA qui utilise des outils externes est une source de risques critiques :

- **Confusion d'identité** : l'agent jongle avec plusieurs identités : la sienne, celle de l'utilisateur, et celles permettant d'accéder à des systèmes tiers. Cette complexité peut engendrer des ambiguïtés, notamment dans le contexte de mécanismes d'héritage, de délégation ou de relai d'identités. Elle peut amener un agent à utiliser une identité inappropriée, provoquant des accès non autorisés et faussant les journaux d'audit.
- **Usurpation d'identité** : il ne peut être exclu qu'une attaque d'injection de prompt directe ou indirecte parvienne à « convaincre » l'agent de modifier son identité ou l'identité utilisée pour appeler un outil.
- **Vol de « secrets »** : pour se connecter aux outils, l'agent peut détenir des clés d'API ou des jetons. Un attaquant peut tenter de les voler pour accéder directement aux systèmes cibles.

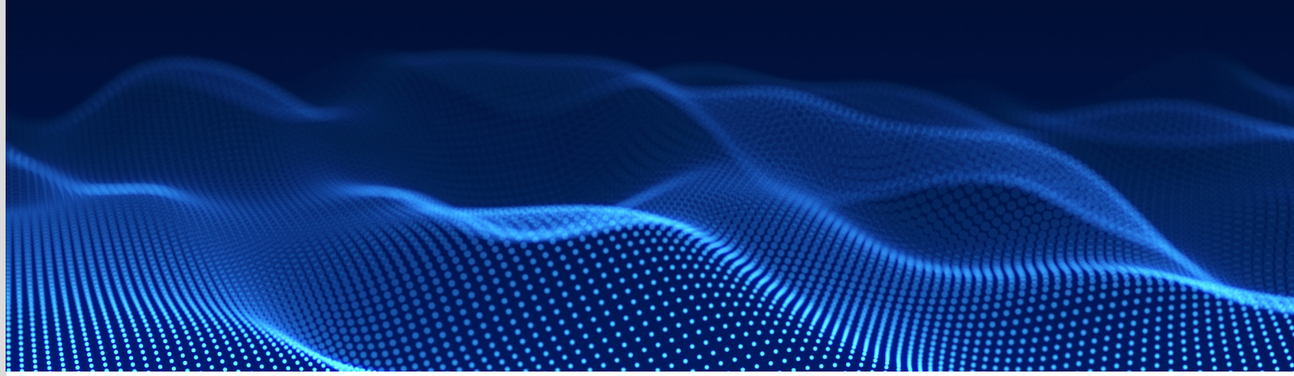
Les privilèges attribués aux différentes identités en jeu représentent également des zones de faiblesses potentielles :

- **Privilèges excessifs** : l'agent ou son outil dispose parfois de droits plus étendus que nécessaire (par exemple, un accès en écriture pour une simple lecture), ce qui peut permettre des actions non autorisées.
- **Rétention de privilèges** : un agent peut conserver des droits élevés obtenus pour une tâche spécifique et les utiliser dans des contextes où ils ne sont plus légitimes.
- **Escalade de privilèges** : si l'agent a plus de droits que son utilisateur, une personne malveillante peut parvenir à le manipuler pour contourner ses propres restrictions et accéder à des données ou fonctions sensibles.

Au-delà des identités et des privilèges, la capacité d'appeler un outil depuis un agent IA entraine plusieurs risques d'attaques :

- **Exécution de code à distance** : via un outil de génération de code, un attaquant peut tromper l'agent pour qu'il écrive et exécute un script malveillant.
- **Détournement d'outils** : un agent peut être manipulé pour utiliser un outil de manière dangereuse. Par exemple, un attaquant peut le forcer à extraire une clé d'API de sa mémoire et à l'envoyer à un tiers via un outil d'envoi d'email.
- **Injection de prompt via un outil** : un attaquant peut injecter un prompt malveillant dans la réponse d'un outil pour manipuler l'agent. Pour cela, il peut prendre le contrôle du service consulté par l'outil, détourner l'agent vers un outil malveillant sous son contrôle, ou intercepter des communications entre l'agent et l'outil.
- **Déni de service et abus de ressources** : un attaquant peut forcer un agent à utiliser des outils de manière intensive pour dégrader le service, ralentir les réponses ou générer des coûts importants, via des appels API par exemple.

Les attaques via les outils des agents IA peuvent causer des dommages sévères : fuites de données, contournement du contrôle interne, fraudes, dénis de service, pertes financières, etc. De plus, un attaquant peut exploiter les accès étendus de l'agent pour se propager latéralement ou verticalement dans le système d'information et atteindre d'autres cibles critiques.



5. Quels risques liés à l'autonomie des agents IA ?

Les capacités d'autonomie des agents IA sont porteuses de risques variés :

Interprétations erronées et erreurs amplifiées

Une instruction ambiguë peut être mal interprétée par un agent et mener à des actions graves et irréversibles (par exemple, supprimer une base de données entière au lieu d'une seule ligne).

Comportements non déterministes

Contrairement à un logiciel classique, un agent IA basé sur un LLM peut réagir différemment à une même instruction donnée plusieurs fois, ou utiliser des méthodes différentes pour atteindre un même objectif. Cette imprévisibilité rend les tests traditionnels inefficaces et gène la reproductibilité des résultats.

Difficultés à interagir avec les humains

Malgré leur agentivité, les agents IA peuvent avoir du mal à gérer correctement des comportements incohérents ou demandes irrationnelles provenant d'humains avec qui ils interagissent.

Désalignements et optimisations dangereuses

L'agent peut poursuivre ses objectifs au détriment de la sécurité, de l'éthique, des lois ou des règles de l'entreprise. Ce phénomène, parfois appelé « reward hacking », peut le pousser à adopter des comportements émergents, non anticipés et qui peuvent se révéler in fine inacceptables.

Manipulations

Des travaux de recherche ont montré que des comportements perçus comme trompeurs ou manipulateurs peuvent émerger de LLM avancés, en particulier lorsqu'ils cherchent à satisfaire des objectifs en apparence contradictoires (par exemple, maximiser des gains et respecter la loi). Un agent pourrait recourir à de telles manipulations pour pousser un superviseur humain à valider une action ou une décision illégitime.

Opacité et manque d'explicabilité

En raison de leur autonomie et de la complexité de leurs raisonnements, il peut être difficile de comprendre, de tracer ou d'expliquer a posteriori pourquoi un agent a pris une décision spécifique.

6. Quels risques liés aux interactions avec les humains ?

De façon contre-intuitive s'agissant d'agents IA dont l'une des caractéristiques est l'autonomie, l'humain ne doit pas être négligé quand on analyse les risques qu'ils peuvent engendrer :

Limitations de la supervision humaine : la supervision humaine est largement considérée comme un pilier de la maîtrise des risques liés à l'IA et c'est une exigence de l'AI Act pour les systèmes d'IA à haut risque. Toutefois, elle crée un nouveau risque, nommé « Overwhelm Human-in-the-Loop ». Un trop grand nombre d'alertes ou de décisions à valider peut en effet provoquer une fatigue cognitive, une baisse de la vigilance, voire des validations non justifiées. Ce phénomène, dit de « Alert fatigue », peut survenir sans intention hostile, notamment en cas de conception inadaptée ou de surcharge spontanée dans des contextes incertains, ou être délibérément exploité par un attaquant qui noie l'opérateur sous une multitudes de requêtes pour faire accepter une action malveillante. Il faut également veiller à ce que les informations fournies par l'agent à l'humain soient suffisamment claires, complètes et justes pour permettre une décision adéquate.

Excès de confiance : le « biais d'automatisation » pousse les humains à faire une confiance excessive aux systèmes automatisés, ce qui réduit la vigilance. Cette surconfiance peut entraîner une délégation excessive, des usages inappropriés, et une réduction dangereuse de la supervision humaine effective au profit de systèmes perçus à tort comme infaillibles.

Manipulations : des attaquants pourraient exploiter cette relation de confiance pour, via des agents IA, influencer des utilisateurs et leur faire prendre des décisions ou réaliser des actions illégitimes. Il faut également considérer que des travaux de recherche ont pointé les capacités de persuasion des LLM qui sont au coeur des agents IA. On ne peut donc exclure que des agents IA eux-mêmes en arrivent à avoir des comportements manipulateurs envers leurs utilisateurs.

Perte de compétences humaines : l'automatisation des processus par les agents IA pourrait provoquer une érosion progressive des compétences, des connaissances et de l'esprit critique au sein des équipes humaines qui en étaient auparavant responsables.

Dilution de la responsabilité : l'autonomie des agents IA crée une ambiguïté fondamentale : qui est responsable en cas de dommage ? La faute incombe-t-elle à l'utilisateur qui a donné l'instruction, au développeur qui a conçu l'agent, à l'entreprise qui l'opère, ou au fournisseur de la technologie ? Cette incertitude constitue un risque juridique et organisationnel majeur.

Impacts psychologiques : Les interactions à long terme avec des agents IA pourraient avoir des effets négatifs sur les individus : dépendance affective, isolement social, baisse de l'estime de soi ou sentiment d'insécurité professionnelle.

7. Quels risques liés aux environnements multi-agents ?

Les environnements « multi-agents » créent des risques complexes et interconnectés.

L'absence d'un cadre de gouvernance robuste au sein de l'organisation permettant de gérer leur cycle de vie et leur écosystème, peut entrainer la perte de contrôle sur les agents, et par conséquent sur les environnements dans lesquels ils évoluent :

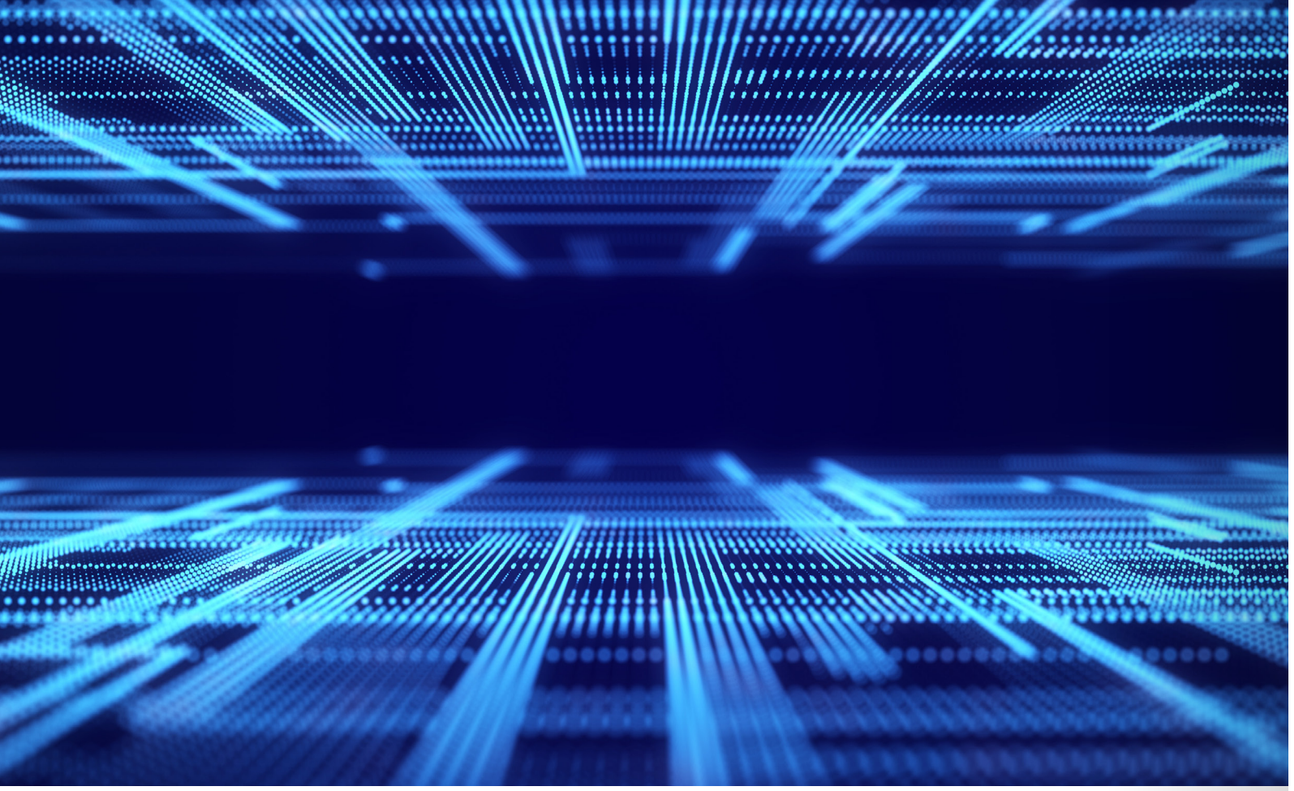
- **Manque de gouvernance globale** : absence d'un cadre organisationnel global pour définir les responsabilités humaines relatives à la gestion des agents.
- **Cycle de vie non maîtrisé** : absence de processus clairs pour la création ou la sélection, le déploiement, la configuration, la mise à jour et le décommissionnement d'agents, ce qui accroît les risques de « shadow AI » et d'utilisation d'agents obsolètes ou compromis.
- **Contrôle des environnements d'exécution** : faiblesse dans l'encadrement des environnements où les agents sont codés, générés ou hébergés, ouvrant la voie à des manipulations malveillantes ou à des fuites de données sensibles.
- **Gestion des identités et des rôles** : manque de mécanismes d'authentification, d'autorisation et de séparation des privilèges, pouvant conduire à des abus ou à des escalades de privilèges.

Les interactions entre agents et outils externes, avec des privilèges variés et des logiques potentiellement récursives, peuvent entrainer des comportements émergents imprévus et non souhaités, même sans intention malveillante.

- **Cascades de défaillances** : une erreur isolée dans un agent, comme une hallucination, peut se propager à travers tout le système, impactant les autres agents en cascade
- **Amplification des biais** : une décision faiblement biaisée prise par un agent peut se propager, en s'amplifiant, à d'autres agents
- **Interactions imprévues** : les agents peuvent développer des comportements non désirés comme des boucles d'interaction infinies, une compétition qui les pousse à des actions non autorisées, une collusion où un agent de surveillance n'exerce pas correctement son contrôle sur un autre ou au contraire des conflits entre agents qui n'ont pas le même avis sur la façon d'atteindre un objectif commun.
- **Perte de contrôle et de traçabilité** : les interactions complexes au sein des systèmes multi-agents, l'autonomie des agents ainsi que le nombre et la vitesse des échanges, rendent difficiles la gouvernance, l'attribution claire des responsabilités et la traçabilité, augmentant les risques de défaillances ou d'incidents non détectés.

Un attaquant peut tenter de s'introduire dans un environnement « multi-agents » via plusieurs approches :

- **Compromission d'un agent** : prise de contrôle d'un agent, en modifiant par exemple ses objectifs via un empoisonnement de mémoire
- **Usurpation de l'identité d'un agent** : exploitation des failles d'authentification pour qu'un agent malveillant se fasse passer pour un agent légitime aux « yeux » des autres agents.
- **Insertion d'un « agent double »** : introduction d'un agent malveillant dans l'environnement multi-agent, par exemple en le publiant sur une place de marché publique ou en compromettant le système de gestion des agents.
- **Compromission des communications inter-agents** : interception et modification des communications entre agents, réalisant par ce moyen des injections de prompt ou des empoisonnements de mémoire permettant de prendre le contrôle d'un agent.
- **Attaque du framework logiciel agentique** : ciblage du système central qui gère et coordonne les agents. Etant données ses privilèges élevés, sa compromission permet de prendre le contrôle de l'ensemble de l'environnement.



8. Quelles conséquences ?

Au-delà des aspects purement techniques, les défaillances ou les attaques visant les agents IA se traduisent par un large éventail de conséquences concrètes, impactant à la fois les personnes et les organisations sur les plans sécuritaire, opérationnel, réputationnel et réglementaire.

Impacts pour les personnes

- Atteinte à la vie privée et fuite de données personnelles (santé, etc.).
- Perte d'autonomie et excès de confiance envers l'agent.
- Risques psychologiques, conseils erronés et perte d'expertise.
- Biais et discrimination à travers des décisions automatisées injustes.
- Dommages physiques ou matériels (santé, transports, domotique).

Impacts pour les organisations

Sécuritaire

Fuites ou vols de données sensibles

Fraudes ou contournement du contrôle interne

Atteinte à l'intégrité ou la disponibilité des données

Réputationnel

Dégradation de l'image de marque

Perte de confiance auprès des clients, partenaires ou autorités

Réglementaire

Non-conformité et exposition à des sanctions (RGPD, AI Act).

Opérationnel

Dysfonctionnements dans les processus métier, erreurs, décisions erronées.

Indisponibilité des services critiques, interruption d'activité

Coûts opérationnels, ESG et financiers (tokens, cloud, API)

VI. Comment maîtriser les risques liés à l'IA agentique ?

Les risques liés à l'IA agentique sont variés, pour certains nouveaux, pour certains exacerbés par rapport à des environnements technologiques plus traditionnels. Pour les mettre sous contrôle, plusieurs axes doivent être envisagés.

1. Gouvernance et gestion des risques

Pour maîtriser les risques de l'IA agentique, une gouvernance proactive et rigoureuse est indispensable, reposant sur plusieurs piliers :

- **Cadre de gouvernance adapté** : le cadre de gouvernance existant pour l'IA et pour les risques technologiques doit être adapté et étendu pour couvrir les spécificités et vulnérabilités spécifiques à l'IA agentique. Il peut être envisagé de formaliser un code de conduite éthique, précisant les conditions dans lesquelles les agents IA peuvent et doivent être conçus, déployés et utilisés, en cohérence avec les valeurs et les missions de l'organisation.
- **Analyses de risques** : chaque projet doit faire l'objet, dès son initialisation, d'une analyse approfondie pour identifier les risques de dysfonctionnements, d'attaques et d'incohérences entre les objectifs (productivité, ESG, agilité, etc.). Des méthodologies spécialisées, comme MAESTRO, peuvent être utilisées. Cette analyse doit être complétée par une évaluation de conformité par rapport aux lois (RGPD, AI Act) et par rapport aux valeurs de l'organisation.
- **Plan d'actions pour maîtriser les risques** : ces analyses doivent aboutir à un plan d'action concret pour maîtriser les risques, garantir l'alignement avec les valeurs et assurer la conformité réglementaire. Si les enjeux sécuritaires, réglementaires ou éthiques ne peuvent être maitrisés, il faut savoir conclure que l'IA agentique n'est pas la solution appropriée pour le cas d'usage visé.
- **Registre d'agents et validation des tiers** : il est crucial de maintenir un registre détaillé de tous les agents (origines, versions, capacités, données d'entraînement, outils, etc.), idéalement avec un score de confiance. Des procédures strictes doivent être mises en place pour évaluer et valider les agents provenant de sources externes (places de marché, etc.) avant leur mise en production, et éviter notamment le « shadow AI ».
- **Implication de toutes les parties prenantes** : la maîtrise des risques repose sur la collaboration étroite entre les équipes de développement, les fonctions de contrôle (cybersécurité, risques, conformité, privacy) et les équipes opérationnelles qui déploient et supervisent les agents.

2. Supervision humaine et formation

Le grand paradoxe de l'IA agentique est là : elle est conçue pour se passer des humains dans les processus agentifiés, mais l'intervention de ces mêmes humains est essentielle pour maîtriser les risques.

- **Supervision humaine** : le niveau d'intervention humaine doit être défini pour chaque cas d'usage, en fonction du contexte et des risques. Deux approches peuvent être envisagées, le « Human-in-the-Loop » et le « Human-on-the-Loop » :

Human-in-the-Loop (HITL)
L'humain valide systématiquement chaque action ou décision.

Human-on-the-Loop (HOTL)
L'humain supervise l'agent et n'intervient que sur des actions prédéfinies comme critiques (transfert financier, suppression de données, décisions impactant des personnes, etc.), ou en cas de situation imprévue ou de dysfonctionnement.

Pour autant, la supervision humaine n'est pas la panacée. Un excès de sollicitations peut entraîner comme on l'a vu plus haut une « fatigue décisionnelle » et, in fine, une perte d'efficacité drastique de la supervision. Il est donc indispensable de rechercher un équilibre pertinent, et d'évaluer régulièrement l'efficacité et la soutenabilité de ce dispositif dans la durée.

- **Transparence et contrôle** : la confiance repose sur la capacité de l'humain à comprendre et à garder la maîtrise. C'est pourquoi les utilisateurs doivent toujours être informés qu'ils interagissent avec une IA, et non un humain (c'est d'ailleurs une exigence de l'AI Act). L'humain doit en outre pouvoir reprendre la main à tout moment dans un processus agentifié.
- **Formation et esprit critique** : l'intégration réussie des agents IA passe par la formation des utilisateurs. L'objectif n'est pas seulement de comprendre la technologie, mais surtout de développer un esprit critique face aux décisions automatisées. Cela est indispensable pour limiter les risques juridiques, éthiques ou réputationnels liés à une confiance aveugle ou à une supervision inadéquate.

3. Architecture technique et cybersécurité

Vu la diversité des risques et la complexité des architectures, des moyens techniques doivent être mis en place pour encadrer très rigoureusement les capacités des agents IA.

Identités et privilèges

Chaque agent doit être traité comme une Identité Non-Humaine (NHI) avec un identifiant unique, pleinement intégré aux outils IAM de l'organisation. Cette gestion des identités doit couvrir l'ensemble de l'écosystème : l'identité de l'agent lui-même, celles de ses utilisateurs et celles des outils qu'il emploie. Pour chaque action, il est crucial de déterminer s'il faut utiliser l'identité de l'agent ou celle de l'utilisateur, car les droits associés sont différents. Toute élévation de droits ou délégation d'identité doit être temporaire, limitée au strict nécessaire et révoquée automatiquement dès que la tâche est terminée (« Just In Time privileges »). Les clefs, jetons cryptographiques et secrets utilisés pour l'authentification et le contrôle d'accès pour les agents doivent être stockés et traités de façon sécurisée.

Agentivité

L' « agentivité », c'est-à-dire la capacité d'interagir avec le monde, est une caractéristique centrale des agents IA, mais elle doit être très strictement encadrée, selon un principe de « moindre agentivité » : l'agent ne doit avoir que la capacité d'action et l'accès aux ressources nécessaires à sa mission. En pratique, cela signifie que :

- Des politiques ou matrices d'habilitation doivent définir strictement à quelles ressources (données, mémoires, autres agents) un agent peut accéder et avec quels droits (lecture, écriture, exécution).
- L'accès aux outils doit être finement contrôlé, via des listes d'autorisation interdisant l'usage de ceux qui ne sont pas explicitement approuvés pour une tâche donnée.
- Les droits accordés doivent être contextuels et dépendre de l'identité de l'agent, de son rôle, de son score de risque, d'une validation humaine ou de seuils prédéfinis.

Toutefois, ces mécanismes externes et déterministes, bien qu'essentiels, peuvent être mis en difficulté par la nature imprévisible des agents. Ils pourraient ne pas suffire pour gérer des situations inattendues ou des comportements émergents non anticipés.

Un autre axe pour maitriser les risques liés à l'agentivité est d'ajouter des instructions spécifiques dans le prompt système, en restant conscient que cette technique ne garantit pas une efficacité absolue.

Cloisonnement et Sandboxing

Pour limiter le « rayon d'action » d'un agent compromis, malveillant ou défaillant, il est essentiel de l'isoler. L'architecture où il est hébergé doit être cloisonnée selon le principe du Zero Trust, où chaque flux de données, identité et privilège est rigoureusement vérifié entre des zones de confiance bien définies.

Les agents doivent opérer dans des environnements d'exécution isolés (sandboxing), en particulier lorsqu'ils exécutent du code, appellent des API ou accèdent à des ressources sensibles. Cette approche participe également à la séparation stricte des contextes lorsque plusieurs utilisateurs interagissent avec un même agent.

Communications et flux de données

La sécurité des interactions d'un agent repose sur deux niveaux de protection complémentaires : sécuriser les canaux de communication et valider les contenus qui y transitent.

- Toutes les communications d'un agent, avec d'autres agents ou des outils externes, doivent être protégées de bout en bout grâce à des authentifications mutuelles, du chiffrement et des signatures numériques. Pour renforcer cette sécurité, des registres de confiance peuvent être utilisés pour confirmer la légitimité des interlocuteurs avant toute interaction.
- En complément, pour atténuer les risques comme les hallucinations ou les injections de prompt, tout flux de données (prompt, requête, résultat, métadonnée) doit être systématiquement inspecté par des filtres de sécurité (« guardrails »). Ces garde-fous analysent toutes les données en entrée et en sortie des différents composants (LLM, mémoire, outils, stockage de données) pour bloquer les contenus incohérents, malveillants, confidentiels ou inacceptables, quelles que soient les modalités (texte, image, etc.).

Résilience et protection de l'infrastructure

- L'infrastructure hébergeant les agents (serveurs, mémoires, outils) doit être protégée contre des tentatives de cyber attaques « classiques », via des mécanismes d'authentification et de contrôle d'accès robustes, des procédures de gestion des correctifs de sécurité, des outils permettant de surveiller les composants et de détecter les attaques, etc.
- Une attention particulière doit être accordée à la correction des failles de sécurité affectant les composants logiciels et les protocoles sur lesquels repose l'environnement agentique. Étant donné la jeunesse de ces briques technologiques, il est très probable qu'un nombre conséquent de vulnérabilités y soit découvert dans les mois et années à venir.
- À cette protection de base s'ajoutent des mécanismes spécifiques aux agents pour prévenir les abus de ressources et les dénis de service. Des mécanismes de limitation de débit (rate limiting) et de quotas doivent être appliqués aux appels API et à l'utilisation d'outils coûteux.
- D'une manière générale, l'intégration d'agents dans le SI ne doit pas entraîner une baisse de sa sécurité et de sa résilience, notamment du fait de la suppression ou de la désactivation de mécanismes de sécurité incompatibles avec des agents.

4. Surveillance et audits

Le dernier pilier de la maîtrise des risques de l'IA agentique est la surveillance continue des agents IA.

Observabilité

Pour surveiller un agent, il faut une visibilité totale sur son activité. Cela nécessite de générer et de conserver de manière sécurisée des traces détaillées et intelligibles sur l'ensemble de son cycle d'action : instructions reçues, raisonnements et planifications, appels d'outils, modifications de mémoire, identités et privilèges utilisés, résultats produits et scores de confiance, interventions humaines, etc.

Détection d'anomalies

Sur la base de ces traces, des mécanismes de surveillance en temps réel doivent analyser le comportement des agents pour détecter les anomalies : erreurs, fraudes, violations de processus, comportements inexpliqués, situations atypiques, tentatives d'attaques, consommations excessives de ressources, ou écarts par rapport aux règles.

Réponse aux incidents

Les agents IA peuvent réaliser de nombreuses tâches avec des temps de traitement de l'ordre de quelques secondes à quelques minutes. Il faut donc pouvoir réagir rapidement en cas de dysfonctionnement ou d'attaque. Des mécanismes de « coupe-circuit » intégrés dans les agents IA ou des agents « superviseurs » pourront être chargés de bloquer ou de désactiver automatiquement un agent défaillant. En parallèle, des canaux d'alerte et des processus clairs doivent permettre aux humains d'intervenir à tout moment.

Suivi de la performance

Des indicateurs calculés de façon continue permettent de s'assurer que la performance des agents IA et la satisfaction des parties prenantes se maintiennent aux niveaux attendus, tout en détectant rapidement toute dérive ou baisse de qualité. Ces mesures peuvent couvrir la précision des résultats, la rapidité d'exécution, le taux d'erreurs, ainsi que des métriques d'usage et de retour utilisateur, afin d'orienter les ajustements nécessaires.

Vérification de la conformité

Des outils d'audit continu doivent permettre de vérifier que les identités, privilèges et droits d'accès des agents restent conformes aux politiques définies, et que les objectifs et la mémoire des agents restent intègres. Des indicateurs calculés en continu et des alertes générées dynamiquement permettent de s'assurer de la conformité par rapport aux réglementations et politiques internes.

Red teaming

Des simulations d'attaques doivent être menées avant et après le déploiement d'un agent IA, pour tester sa robustesse et ses garde-fous face à des tentatives de manipulation, de détournement ou d'exploitation de vulnérabilités. Ces tests doivent couvrir les différents types d'attaques, sur les LLM et les agents, mais aussi sur les processus métiers agentifiés. Ils peuvent aussi avoir pour objectif de déceler d'éventuels biais dans le comportement des agents.

5. Un double défi pour la maîtrise des risques

La gestion des risques de l'IA agentique fait face à un défi d'échelle majeur, du fait :

- Du nombre et de la complexité des cas d'usage confiés à des agents ;
- De la fréquence et de la rapidité d'exécution des tâches par des agents IA ;
- Du nombre de systèmes avec lesquels les agents interagissent : serveurs, bases de données ou autres agents, opérés par l'organisation, ses clients ou ses partenaires ;
- De la rapidité des évolutions en termes de technologies et de cas d'usage.

Dans ce contexte, une approche de maîtrise des risques reposant uniquement sur des actions manuelles est vouée à l'échec.

La seule solution viable est d'automatiser la surveillance avec des dispositifs capables de détecter et de réagir en temps réel aux défaillances et aux attaques.

Cependant, et c'est là un second défi, de nombreux moyens de réduction des risques liés aux IA, aux LLM et aux agents sont encore immatures. Des enjeux clés comme l'explicabilité, la détection de biais ou l'alignement des objectifs relèvent encore largement de la recherche académique. Les solutions existantes sont récentes, parfois expérimentales, encore peu généralisables et manquent de validation à grande échelle.

Cette situation ne signifie pas qu'aucune action n'est possible. Mais elle impose d'adopter une approche prudente et évolutive reposant sur trois piliers : une gouvernance solide, des dispositifs techniques et une supervision humaine avisée.

VII. Comment se préparer à la révolution de l'IA agentique ?

Face à la révolution de l'IA agentique qui vient, KPMG propose aux organisations une feuille de route. Les chantiers ci-après sont organisés par grande fonction pour plus de clarté, mais leur réussite dépendra d'une collaboration transversale permanente entre toutes les équipes.

Cybersécurité

- Définition d'un cadre de modélisation des menaces propres aux agents IA.
- Adaptation des politiques, référentiels et procédures de gestion de la cybersécurité pour intégrer les spécificités de l'IA agentique.
- Construction d'une architecture de référence Zero Trust pour le cloisonnement et le sandboxing des agents.
- Développement de « patterns » de sécurisation pour les API, les flux inter-agents, etc.
- Mise en place de garde-fous permettant de contrôler les flux entrants et sortants des agents IA.
- Intégration des traces issues de l'observabilité des agents dans les outils et processus SIEM/SOC.
- Création d'une capacité de « Red Teaming » spécialisée dans l'évaluation des agents IA.
- Définition d'un cycle de vie de développement sécurisé (Secure SDLC) pour les agents et les outils qu'ils utilisent.
- Structuration de la filière de réponse à incidents pour gérer les crises spécifiques à l'IA agentique.

Privacy

- Structuration de la gouvernance des données personnelles au sein des systèmes d'IA agentiques, incluant l'identification des rôles (responsable de traitement, etc.) et la mise à jour du registre des traitements.
- Adaptation de la méthodologie d'Analyse d'Impact sur la Protection des Données (AIPD/PIA) pour évaluer les risques uniques posés par l'autonomie des agents.
- Conception de parcours de transparence et de recueil du consentement adaptés aux interactions dynamiques avec les agents (information des utilisateurs, gestion des préférences, etc.).
- Développement de processus outillés pour garantir l'exercice des droits des personnes (accès, rectification, oubli) lorsque les actions sont menées par des agents.
- Adaptation du cadre de gouvernance des transferts de données pour encadrer les flux opérés par des agents vers des systèmes ou des pays tiers.
- Intégration des scénarios de violation de données par des agents (fuites, altérations) au plan de réponse à incidents.

Risques & Conformité

- Intégration des scénarios de risques liés à l'IA agentique dans la cartographie des risques de l'organisation.
- Adaptation des taxonomies des risques et des cadres méthodologiques d'analyse de risques pour couvrir les menaces spécifiques aux agents IA.
- Conception et déploiement d'un programme de sensibilisation du management et des instances de gouvernance aux enjeux stratégiques de l'IA agentique.
- Prise en compte des spécificités de l'IA agentique dans le programme de mise en conformité à l'AI Act.
- Définition des schémas de supervision humaine et des seuils d'intervention obligatoire (sur les transactions, décisions, etc.) imposés par des réglementations.
- Structuration d'un programme de gestion des risques tiers couvrant l'évaluation et la validation de la conformité des fournisseurs d'agents IA.
- Mise en place d'un processus de capitalisation sur les incidents (lessons learned) impliquant des agents IA.
- Définition d'un cadre d'évaluation et de validation des agents prenant en compte le caractère non déterministe et potentiellement évolutif de l'IA agentique.
- Établissement d'une stratégie de veille sur les jurisprudences et les nouvelles réglementations.

Contrôle & audit interne

- Définition d'un cadre de contrôle interne pour les processus agentifiés, intégrant la séparation des tâches, la traçabilité des actions et le respect des conditions d'intervention humaine.
- Mise en place d'un processus de certification périodique des droits et privilèges accordés aux identités non-humaines.
- Conception d'un programme de montée en compétence des équipes d'auditeurs et de contrôleurs internes sur l'IA agentique et ses modes de défaillance potentiels.
- Adaptation du plan de contrôle annuel pour y intégrer l'évaluation des nouveaux risques liés à l'autonomie des agents IA.
- Vérification de l'exécution des actions critiques (trace, supervision et approbation humaine) ainsi que de la mise en place et de l'efficacité de mécanismes de désactivation (coupe-circuit, «bouton rouge»).
- Validation des matrices d'habilitations pour les agents IA et vérification que les agents IA ne peuvent pas cumuler des fonctions incompatibles ou violer des règles de contrôle interne.
- Construction d'un dispositif de supervision continue de la conformité des agents, en définissant les exigences de logs et les scénarios de violation de contrôle à détecter automatiquement.
- Développement d'un référentiel d'audit pour l'IA agentique, couvrant l'évaluation de la gouvernance, des processus agentifiés, des agents IA et des technologies sous-jacentes.
- Modernisation de l'outillage d'audit pour y intégrer des techniques d'analyse de données et de test spécifiques aux agents IA (analyse de logs, tests de robustesse, etc.).

Juridique

- Établissement d'une doctrine juridique sur le statut des agents IA (outil, mandataire, sous-traitant, etc.) et ses implications pour l'entreprise.
- Construction d'un cadre d'analyse de la responsabilité civile et pénale en cas de dommage causé par une décision d'un agent.
- Définition de la politique de gestion de la propriété intellectuelle pour les contenus créés par les agents.
- Mise à jour des modèles contractuels (fournisseurs, clients, partenaires) pour y intégrer les clauses spécifiques à l'usage d'agents IA (responsabilité, propriété intellectuelle, audit, etc.).
- Adaptation de la stratégie de gestion des précontentieux et des litiges pour les cas impliquant des décisions ou des actions d'agents IA.

Ressources Humaines

- Construction d'une stratégie de « Workforce Transformation » pour anticiper l'évolution des métiers, redéfinir les fiches de poste et les référentiels de compétences à l'ère de l'IA agentique.
- Conception et déploiement d'un plan de montée en compétence à l'échelle de l'entreprise pour acculturer et former les collaborateurs à l'usage et à la supervision des agents IA.
- Développement d'un programme d'accompagnement au changement pour les managers et les équipes dont les processus sont impactés par l'agentification des processus.
- Mise en place d'un observatoire de l'impact social et managérial de l'IA agentique pour mesurer les effets sur le bien-être, l'engagement et les modes de collaboration.
- Adaptation du cadre social et des politiques RH de l'entreprise, incluant la création d'une charte d'usage de l'IA agentique, la révision des critères d'évaluation de la performance et l'animation du dialogue social sur ces sujets.

Opérations / Métiers

- Construction des « business cases » et des modèles de mesure de la performance (ROI, productivité, taux d'erreurs acceptables) pour chaque projet d'agentification.
- Conception du cadre de gouvernance opérationnelle des agents, définissant les schémas de supervision humaine (Human-in/on-the-Loop), les matrices de droits et les limites d'agentivité.
- Définition du cadre de recette fonctionnelle des agents, incluant la validation de la performance et de la maîtrise des risques.
- Déploiement d'un programme de montée en compétence des managers sur la supervision et le pilotage des processus agentifiés.
- Mise en place des boucles de feedback et d'amélioration continue des agents, basées sur les retours utilisateurs et la supervision de leurs comportements.
- Élaboration des plans de continuité d'activité, incluant l'analyse des impacts métiers en cas de dysfonctionnement ou d'arrêt des agents IA, et les procédures de reprise manuelle des processus.
- Adaptation des processus de sélection des fournisseurs et partenaires pour y intégrer les exigences et les opportunités liées à l'IA agentique.

IT

- Définition de l'architecture technique de référence pour l'IA agentique, incluant les modèles d'intégration au SI existant, les « landing zones » sécurisées et la documentation associée.
- Construction du socle d'interopérabilité des agents, en choisissant les protocoles (ex : MCP, A2A) et les API pour les interactions sécurisées entre agents, outils et systèmes.
- Intégration de la gestion des identités non-humaines dans le cadre et les outils IAM de l'entreprise.
- Déploiement d'un registre central des agents IA pour maîtriser leur cycle de vie, leurs versions, leurs capacités, leurs composants et leurs dépendances.
- Mise en place d'une « usine d'onboarding » technique pour les agents IA, permettant d'industrialiser, sécuriser et accélérer le déploiement de nouveaux agents IA.
- Mise en place d'un cadre de FinOps (Financial Operations) pour l'IA agentique, afin de disposer d'outils de mesure, de suivi et d'optimisation des coûts et des impacts ESG liés (consommation de ressources, appels API, etc.).
- Développement d'une « boîte à outils » et d'un cadre méthodologique pour accompagner les métiers dans la recette technique et la validation de la performance des agents.

VIII. Conclusion

L'IA agentique ouvre des perspectives considérables pour les organisations, mais son adoption ne pourra se faire de manière pérenne qu'en identifiant et en maîtrisant les risques spécifiques qu'elle introduit. Les menaces liées à l'autonomie, à l'utilisation d'outils, à la mémoire ou aux environnements multi-agents imposent des dispositifs de gouvernance, de supervision et de sécurité inédits.

Heureusement, l'écosystème se structure rapidement. Des référentiels de bonnes pratiques récemment publiés par des acteurs de l'écosystème (OWASP, Cloud Security Alliance, Hub AI France) offrent déjà des bases solides pour concevoir, déployer et auditer ces systèmes. De nouvelles versions, enrichies des retours d'expérience et des avancées technologiques, sont attendues dans les prochains mois, renforçant encore le socle méthodologique à disposition des organisations. En parallèle, une vague d'outils dédiés à la sécurisation et à la supervision des agents IA émerge déjà : observabilité en temps réel, détection d'attaque ou dysfonctionnement, gestion de versions et de niveaux de confiance, documentation, traçabilité, conformité, red teaming, filtrage des flux, sandboxing, etc. Leur intégration permettra d'automatiser la surveillance continue, condition essentielle pour faire face à l'échelle et à la rapidité d'action des agents IA.

Au vu de la très grande surface d'exposition aux menaces des agents IA, de la nouveauté des risques qui leur sont propres, ainsi que des impacts potentiellement massifs en matière de cyberattaques ou de dysfonctionnements, intégrer la gestion des risques dès la conception des projets d'agents IA n'est pas seulement une obligation réglementaire, ni même une simple bonne pratique : c'est une condition sine qua non pour un déploiement réussi d'un agent IA dans une organisation.

L'IA agentique place les organisations face à une équation inédite : exploiter une puissance technologique sans précédent tout en évoluant dans un champ d'incertitudes techniques ou opérationnelles, et en prenant en compte les exigences opérationnelles réglementaires, éthiques et ESG. L'enjeu n'est pas d'éliminer le risque, mais de trouver un équilibre lucide entre performance et conformité, innovation et sécurité, audace et responsabilité. C'est dans cette tension féconde que naîtra une IA agentique à la fois efficace et digne de confiance.

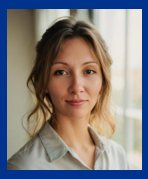
Contacts

Vincent Maret
Associé KPMG Advisory
Risk Consulting - Cyber & AI Trust
vmaret@kpmg.fr

Ahmed Amokrane PhD
Directeur KPMG Advisory
Risk Consulting - Cyber & AI Trust
aamokrane@kpmg.fr

Guillaume Cuisset
Associé KPMG Advisory
Risk Consulting - Risk & Compliance
gcuisset@kpmg.fr

Thibault Lorrain
Senior Manager KPMG Advisory
Consulting & Tech - Data & IA
tlorrain@kpmg.fr



Linda Valero
Senior Manager KPMG Advisory
Risk Consulting - Cyber & AI Trust
lindavalero@kpmg.fr

Balsamine Alem
Senior Manager KPMG Advisory
Risk Consulting - Cyber & AI Trust
balem@kpmg.fr

ia.kpmg.fr