HUB FRANCE IA

NOTICE

GOUVERNANCE DES AGENTS EXPERTS D'IA GENERATIVE

Juillet 2025

Table des matières

Introduction

La gouvernance des agents d'intelligence artificielle (IA) est devenue un enjeu crucial pour les organisations souhaitant intégrer ces technologies de manière éthique et responsable, tout en conciliant l'agilité nécessaire à leur adoption à grande échelle et à la maîtrise des nouveaux risques. La présente notice examine en détail les objectifs, les responsabilités, les acteurs et les mesures de performance associés à la gouvernance de ces agents.

L'émergence des agents experts d'intelligence artificielle générative (agents *GenAI)* incite les organisations à adapter leur gouvernance existante. En effet, les organisations anticipent une démocratisation de l'usage des agents IA connectés aux données d'entreprise et manipulant des outils pour réaliser des actions automatisées[1]. Ces agents d'IA sont désormais encadrés par l'*Artificial Intelligence Act (AI Act)*[2], qui rend obligatoire une analyse des risques inhérents. Par conséquent, l'évolution de la gouvernance des entreprises due à l'*AI Act* sera fortement influencée par la démocratisation des agents *GenAI*.

Cette évolution soulève de nouvelles préoccupations liées à l'utilisation des agents d'IA, telles que :

- **Le partage excessif de données :** la gestion des droits d'accès aux données d'entreprise par les organisations est souvent déficiente. L'absence de labellisation de données sensibles et l'attribution excessive de droit d'accès, notamment aux données sensibles, constituent des problèmes majeurs. Désormais, les agents *GenAI* utilisés par des utilisateurs dotés de tels droits excessifs peuvent accéder à l'ensemble de ces informations et les agréger, augmentant ainsi les risques d'utilisation frauduleuse.
- **La fuite de données sensibles via les agents *GenAI* :** les utilisateurs demandent à ces agents de générer des rapports et des synthèses à partir des données internes. Les réponses des agents *GenAI* peuvent inclure des données sensibles

[1] World Economic Forum. Navigating the AI Frontier: A Primer on the Evolution and Impact of AI Agents. White paper. December 2024. https://reports.weforum.org/docs/WEF_Navigating_the_AI_Frontier_2024.pdf
[2] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance) https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32024R1689

dont l'utilisateur ne perçoit pas le niveau de sensibilité. L'utilisateur partage alors largement ces informations sans connaître le niveau de sensibilité.

- **L'utilisation non conforme des agents *GenAI* :** les utilisateurs, qui peuvent également être des développeurs de ces agents, peuvent les utiliser pour des actions non éthiques ou relevant des catégories à haut risque définies par l'*AI Act*.
- **La lenteur de réponses des gouvernances existantes :** La démocratisation de l'utilisation des agents *GenAI* via le *no code*, accentuée par la pression exercée sur les employés pour se différencier et devenir proactifs face à l'adoption de l'IA, ne peut plus se permettre d'attendre les délais de réponse des comités et augmente le risque de *ShadowIT+AI*.
- Par ailleurs, les agents *GenAI*, dont les modèles génératifs sont souvent au cœur du processus décisionnel, constituent des vecteurs supplémentaires de propagation des risques associés aux modèles génératifs[3]. Parmi les principales menaces, on trouve :
  - L'injection de prompt consiste à insérer des commandes malveillantes pour influencer le comportement de l'agent *GenAI*.
  - Le *jailbreak, qui* permet de contourner les restrictions de sécurité afin d'accéder à des fonctionnalités non autorisées.
  - L'empoisonnement des données, qui manipule les données d'entraînement pour fausser les résultats de l'agent *GenAI*.
  - Le détournement de modèle, qui implique la prise de contrôle des modèles d'IA à des fins malveillantes.
  - L'abus de portefeuille, qui exploite les systèmes de paiement intégrés utilisés par les agents *GenAI*.

Il est ainsi crucial pour les organisations de mettre en place des mesures de sécurité robustes afin de protéger leurs systèmes contre ces menaces émergentes et d'adapter leur gouvernance pour devenir à la fois plus agiles et réactives, tout en étant plus prudentes face aux risques multiples et évolutifs. Chaque technique d'IA comporte ses propres risques, dont certains sont encore inconnus et qui se manifesteront bien après leurs premières mises en œuvre, et dont l'impact dépend de l'usage. Dans la première partie, nous présentons de manière non exhaustive les risques liés à l'utilisation de systèmes basés sur des agents *GenAI*.

[3] Voir le Livre blanc : Analyse des attaques sur les systèmes de l'IA, Hub France IA et Campus Cyber, mai 2025 : https://www.hub-franceia.fr/wp-content/uploads/2025/05/25_05_11_Analyse-des-attaques_sur_les_systeme_de_l_IA_VF.pdf

HUB FRANCE IA

1. Les risques des agents GenAI

1. Les risques des agents *GenAI*

Comparativement à l'usage d'un chatbot qui se fait toujours sous la supervision d'un humain, les risques des agents *GenAI* sont amplifiés par deux facteurs clés : la capacité d'exécuter des actions[4] et la capacité de le faire avec plus ou moins de supervision/contrôle humain. Un agent classant automatiquement des documents comptables par exemple n'a pas le même impact qu'un agent *GenAI* chargé d'optimiser les achats en ayant la responsabilité sans supervision de déclencher des demandes d'achat. La gestion des risques autour des agents *GenAI* devient beaucoup plus dynamique et évolutive afin de minimiser l'impact potentiel suivant différentes dérives selon trois niveaux de granularité potentiels à avoir en tête :

- Au niveau du ou des composants d'IA générative utilisés (ex : *Large Language Mode*l, ou IA multimodale). Le Hub France IA a publié en juillet 2024 un livre blanc[5] qui détaille les différents risques des IA génératives, leurs causes et potentiels impacts suivant différentes dimensions (juridiques, financières, opérationnelles, …). Il présente une démarche d'analyse des risques génériques appliquée à différents cas d'usage dans la finance, le marketing, la cybersécurité ou d'autres secteurs d'activité. Enfin ce livre blanc introduit un certain nombre de pistes d'atténuation et de remédiation des risques.
- Au niveau d'un agent *GenAI*, on peut citer[6] :
    - De **potentielles fuites d'informations sensibles**. Les fuites de confidentialité surviennent lorsque les agents, en raison de leurs interactions avec des applications, demandent des informations personnelles sensibles, ce qui augmente le risque d'extraction de données. Les agents *GenAI* doivent gérer les sessions pour maintenir la confidentialité et l'intégrité des données échangées entre les utilisateurs et le serveur. La gestion des sessions et des droits est complexe, car les agents doivent suivre les interactions des utilisateurs dans le temps. Si cette gestion est inadéquate, cela peut entraîner des fuites d'informations et des assignations d'actions erronées.
    - **Des défaillances dans les protocoles de sécurité**. Un agent *GenAI* peut avoir des accès/privilèges plus importants qu'un utilisateur (voire un attaquant) et

[4] Jinwei Hu, Yi Dong , Shuang Ao, Zhuoyun Li, Boxuan Wang, Lokesh Singh, Guangliang Cheng, Sarvapali D. Ramchurn, Xiaowei Huang. Position: Towards a Responsible LLM-empowered Multi-Agent Systems. *arXiv preprint arXiv:2502.01714*. (2025). https://arxiv.org/pdf/2502.01714

[5] Livre blanc. Les risques de l'IA générative. Hub France IA. Juillet 2024. https://www.hub-franceia.fr/wp-content/uploads/2024/09/Hub-France-IA-Les-risques-de-lIA-Generative-final.pdf

[6] La liste n'est pas exhaustive

donner à ce dernier la possibilité d'effectuer des actions non autorisées. Les droits d'accès de l'agent *GenAI* peuvent être attachés à un utilisateur et si ce dernier change de fonction dans une même entreprise sans que ses anciens droits aient été supprimés, cela peut engendrer des failles de sécurité. L'isolation des privilèges, la limitation des droits de l'agent à ceux de l'utilisateur pour lequel il agit et l'identification des requêtes légitimes sont des méthodes permettant de diminuer ce risque de faille dans les protocoles de sécurité, notamment en cas d'attaque basée sur l'injection de prompt.

- **L'empoisonnement de la mémoire** désigne une technique visant à manipuler les systèmes de mémoire à court terme ou à long terme d'un agent. Cette méthode consiste à injecter des données falsifiées ou malveillantes afin d'influencer le contexte et de manipuler l'agent (ex : altération des décisions, actions non autorisées).
- **Une vulnérabilité accrue aux attaques par déni de service** (*DoS*) en raison de la charge computationnelle potentiellement élevée des requêtes et des appels répétés à des outils par l'agent.
- **L'utilisation malintentionnée d'outils** par des cybercriminels, notamment via l'exécution de code à distance (*Remote Code Execution*).
- **Des erreurs dans l'exécution des tâches ou d'appels aux outils adéquats** car l'utilisation d'une IA générative comme composant pour interpréter et générer des instructions introduit éventuellement des imprécisions impactant le fonctionnement des outils (ex : formats attendus des entrées, paramètres manquants dans l'appel de l'outil, …). Un agent pourrait appeler par erreur un outil qui effectuerait une tâche potentiellement dangereuse (ex : écriture dans une base de données par exemple).
- **L'incompréhension des contextes**. Les agents *GenAI* peuvent mal interpréter des contextes ou ne pas disposer d'informations suffisantes, entraînant des actions inappropriées ou dangereuses.

- Au niveau d'un système multi-agent, on peut citer[7] :
  - **La propagation des hallucinations (Dérive de Connaissances et propagation d'erreurs)**. Lorsque les modèles d'IA générative intégrés aux agents *experts* génèrent des informations incorrectes, les erreurs peuvent se propager entre ces agents. Dans les tâches de raisonnement collaboratif, les agents *GenAI* peuvent s'aligner sur un consensus erroné en raison de phénomènes comme la conformité et le biais d'autorité. Par exemple, dans des débats entre agents *GenAI*, un agent expert ayant une compréhension erronée peut générer des

[7] La liste n'est pas exhaustive

raisons persuasives mais fausses, impactant les autres et les détournant des chemins de raisonnement vers des solutions correctes. Ce problème constitue une dérive de connaissances.

- **La propagation de l'incertitude.** À mesure que les systèmes deviennent plus complexes, les incertitudes inhérentes aux agents *GenAI* individuels peuvent s'accumuler, compromettant potentiellement la stabilité globale du système. Les agents *GenAI* montrent également une tendance à l'expansion des biais cognitifs, amplifiant et propageant les erreurs plutôt que de les filtrer, ce qui aggrave la dérive de connaissances.[8] Les solutions existantes, telles que l'ingénierie des prompts et les interventions de type « humain dans la boucle », sont souvent limitées en termes d'évolutivité et de praticité. Une piste pour résoudre ces problèmes consiste à utiliser une architecture intégrant des mécanismes de quantification de l'incertitude dans ses principes opérationnels, garantissant un alignement cohérent des connaissances à travers le réseau d'agents. Il existe plusieurs sources d'incertitudes : **l'incertitude inhérente à chaque agent expert** et **l'incertitude dans les interactions entre Agents *GenAI***. La performance des systèmes multi agents (de type SMA-LLM[9]) peut être évaluée par des métriques statistiques ou grâce à une vérification humaine (de type « humain dans la boucle »). Identifier les résultats en sortie du système ayant un niveau d'incertitude associée élevé pour une évaluation ultérieure par un humain contribue à renforcer la robustesse des systèmes multi-agents, à condition que le nombre d'éléments incertains reste raisonnable et absorbable par les opérateurs humains.
    - **L'intelligence collective**. L'un des défis est d'atteindre une compréhension mutuelle entre agents *GenAI* pour maximiser l'intelligence collective. Contrairement aux SMA traditionnels, qui reposent sur des protocoles prédéfinis, ces agents basés sur des LLM présentent des comportements émergents et imprévisibles en raison de leur fonctionnement et de leur entraînement sur des ensembles de données vastes et variés. Cette imprévisibilité nécessite le développement de mécanismes quantifiables, tels que des métriques de confiance, pour faciliter une coordination efficace entre agents *GenAI*. Sans ces mécanismes, les agents *GenAI* peuvent avoir du mal à interpréter ou à s'aligner sur les actions des autres humains ou agents. La notion de coopération dans les SMA-LLM (**Coopération Agent-Agent**) est essentielle pour garantir la cohérence responsable et opérationnelle. La coopération se manifeste par la capacité des agents *GenAI* à traiter les intentions et les sorties des autres agents. Elle peut reposer sur des capacités de délégation : des

[8] Moshe Glickman, Tali Sharot. "How human–AI feedback loops alter human perceptual, emotional and social judgements". *Nat Hum Behav* 9, 345–359 (2025). https://www.nature.com/articles/s41562-024-02077-2
[9] LLM : *Large Language Model* SMA-LLM : Système multi agents basé sur des *Large Language Models*

modèles avec des capacités de raisonnement plus élevées guident des modèles plus faibles en leur déléguant des tâches. Pour atteindre et évaluer efficacement la coopération globale dans un système multi-agents, des méthodes d'évaluation dédiées sont essentielles[10]. Des métriques telles que les ratios de coopération et de coordination, les scores de confiance et la similarité sémantique sont proposées pour évaluer la qualité de la collaboration entre agents.

- Lorsque la coopération n'est pas souhaitée, il y a un risque de **collusion**. La collusion peut survenir à la fois des communications entre agents et des mécanismes internes des agents individuels.
- Les **conflits entre agents** dans les SMA-LLM proviennent généralement d'un désalignement des objectifs et d'une asymétrie des connaissances. Les conflits au niveau des objectifs peuvent émerger d'interprétations divergentes d'un même objectif de haut niveau, entraînant des stratégies d'exécution divergentes. Les conflits basés sur les connaissances surviennent lorsque les agents *GenAI* construisent des modèles mentaux différents malgré des informations initiales identiques. La nature probabiliste des LLM et les ambiguïtés sémantiques inhérentes amplifient les effets du désalignement des connaissances.
- **La compréhension Agent-Humain**. Pour opérer avec des humains dans la boucle (**coopération Agent-Humain**), les agents *GenAI* doivent interpréter avec précision le langage naturel et le contexte (ex : contraintes sociétales). Des méthodes telles que l'apprentissage par renforcement à partir de retours humains (RLHF[11]), le perfectionnement supervisé (SFT[12]) et l'optimisation des préférences (PO[13]) sont couramment utilisées pour aligner les sorties des agents/LLM sur les valeurs humaines. Cela nécessite de travailler aussi sur la désambiguïsation du langage notamment lorsque le langage est très

[10] Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, Xiangliang Zhang. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*. 2024. https://arxiv.org/pdf/2402.01680

[11] RLHF : *Reinforcement Learning from Human Feedback* (apprentissage par renforcement à partir de rétroaction humaine). Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*. 2019. https://arxiv.org/pdf/1909.08593

[12] SFT : *Supervised Fine-Tuning*. Il consiste à adapter des modèles de langage pré-entraînés à des tâches spécifiques, en les entraînant sur un ensemble de données spécifique à la tâche ou domaine, avec des exemples étiquetés.

[13] PO : Preference Optimization. Cette technique consiste à aligner le modèle de langage grâce à une base de données contenant pour chaque *prompt* une réponse préférée et une non préférée.

spécifique à un métier et donc potentiellement soumis à mésinterprétation par les LLMs entrainés sur des corpus généralistes.

- **Les attaques par empoisonnement de données et le *jailbreaking***. Les attaques par empoisonnement de données et le *jailbreaking* introduisent des vulnérabilités dans les SMA-LLM en exploitant les canaux de communication. La dépendance à des interactions dynamiques et à des outils/connaissances externes élargit les surfaces d'attaque, nécessitant des mécanismes dédiés pour détecter et filtrer les données compromises.
- **Les menaces cyber sur les architectures distribuées.** Les menaces cyber posent également des défis importants aux SMA-LLM en raison de leur architecture distribuée. Par exemple, des attaques au niveau du réseau peuvent perturber les performances.
- ***Le Scaling up*.** Lorsque le nombre d'agents IA dans un système augmente, cela conduit à une augmentation de la complexité calculatoire. Le système a besoin d'une puissance de calcul et de mémoire accrues. L'orchestration des agents devient clé.
- Le potentiel manque de **traçabilité** (effet « boîte noire ») et de **reproductibilité** des décisions dans les SMA-LLM représente un risque à prendre en compte lors de l'utilisation de ce type de système.

HUB FRANCE IA

2. Objectifs de la gouvernance des agents GenAI

2. Objectifs de la gouvernance des agents *GenAI*

2.1. Assurer l'éthique et la conformité

L'un des principaux objectifs de la gouvernance des agents *GenAI* est de garantir que leur utilisation respecte les normes éthiques et les réglementations en vigueur, tout en apportant un réel gain de productivité pour l'entreprise ou la collectivité, sans pour autant nuire à l'environnement. La gouvernance a donc pour rôle de conserver cet équilibre délicat, qui évoluera sans doute, et cela inclut la protection des données personnelles et la transparence dans les processus décisionnels des agents *GenAI*.

L'*AI Act* interdit l'utilisation d'agents d'IA qui présentent un risque inacceptable pour les droits fondamentaux des individus. Ces agents d'IA incluent :

- Les agents d'IA de notation sociale : l'évaluation des individus basée sur leur comportement social, susceptible d'entraîner des discriminations.
- Les agents d'IA utilisant l'identification biométrique en temps réel : l'utilisation des technologies de reconnaissance faciale dans des espaces publics sans consentement.
- Les agents d'IA utilisant la manipulation comportementale : les systèmes qui exploitent les vulnérabilités des individus pour influencer leurs décisions, notamment dans des contextes tels que le recrutement ou l'éducation[14].

Au-delà des interdictions, l'*AI Act* exige que chaque projet d'IA évalue la nature des risques et mette en place des mécanismes de gestion de ces risques. Or dans le cas de l'IA agentique, il sera nécessaire d'intégrer à la gouvernance une validation lors de l'industrialisation, dès le prototype réalisé par un utilisateur non spécialiste. Pour les agents les plus impactants, un suivi dans la durée tout au long de leur utilisation sera également requis, tout en veillant à conserver un bon équilibre vis-à-vis de l'humain. Le récent retour de la FinTech Klarna de remplacement des employés par des agents IA démontre combien cette recherche d'équilibre est complexe[15].

[14] Virgine Dignum. Responsible Artificial Intelligence: Designing AI for Human Values. AI & Society, 36(1), 1-12 2021 https://www.itu.int/dms_pub/itu-s/opb/journal/S-JOURNAL-ICTF.VOL1-2018-1-P01-PDF-E.pdf
[15] Frédéric Charles. Klarna nous montre les limites des agents IA. ZDNet. 11 mai 2025. https://www.zdnet.fr/blogs/green-si/klarna-nous-montre-les-limites-des-agents-ia-474924.htm.

2.2. Promouvoir la transparence

La transparence est essentielle pour instaurer la confiance parmi les parties prenantes. Les organisations doivent s'assurer que les décisions prises par les agents *GenAI* sont compréhensibles et justifiables notamment lorsque ces décisions impactent des humains ou leur organisation. Cela nécessite une documentation claire des critères algorithmiques et des données utilisés : ces informations ne sont pas toujours disponibles (notamment pour les *LLM*), l'*AI Act* prévoit des exigences sur ce point. Il est également crucial de mettre en place des mécanismes de transparence pour informer les utilisateurs sur le fonctionnement des agents *GenAI*, ainsi que d'assurer un accompagnement sociotechnique, notamment via un dialogue social plus ouvert, c'est-à-dire une écoute franche et ouverte des acteurs impactés par l'utilisation de ces agents.

2.3. Assurer la qualité des données

La démocratisation des agents *GenAI* accédant à l'ensemble des données utilisateurs pour effectuer de multiples tâches soulève des enjeux cruciaux de gouvernance de la donnée, notamment en matière de sécurisation des données. Cette sécurisation est essentielle pour limiter les erreurs des agents IA (hallucinations) et la diffusion d'informations sensibles.

Plusieurs actions sont à mener pour atteindre cet objectif :

- **Connaître ses données**. Il est essentiel de comprendre la nature, l'origine et l'utilité des données. Les bonnes pratiques incluent l'inventaire des données, leur classification selon leur sensibilité et l'utilisation d'outils de gestion des métadonnées intégrant le « contexte » de collecte, de traçabilité, de traitement et de filtrage de ces données.
- **Assurer la pérennité des données**. Cela implique de maintenir la pertinence et l'utilisabilité des données tout au long de leur cycle de vie. Les pratiques recommandées incluent la définition de politiques de conservation, la mise à jour et le nettoyage des données, ainsi que l'utilisation de technologies de stockage appropriées.
- **Protéger ses données**. La sécurité des données est primordiale pour prévenir les accès ou les utilisations non autorisés. Les bonnes pratiques incluent la mise en œuvre de mesures de sécurité telles que le chiffrement, la réalisation d'audits de

sécurité réguliers et la sensibilisation des employés aux risques. Il convient également de tenir compte de la propriété des données, notamment celles acquises et donc soumises à une licence d'utilisation, qui peuvent être exposées à des contraintes contractuelles d'utilisation d'agrégation et/ou de revente.

- **Prévenir la perte des données**. Il est crucial de mettre en place des stratégies pour éviter la perte de données, qu'elle soit due à des erreurs humaines ou à des incidents techniques. Cela comprend des systèmes de sauvegarde réguliers, des plans de reprise d'activité et des tests des procédures de récupération.

Certaines organisations, comme La Poste par exemple, envisagent la mise en place de données synthétiques pour réaliser des démonstrations de preuve de concept (*PoC*) sans données sensibles. L'objectif est de créer des jeux de données fictifs mais représentatifs, non traçables jusqu'aux données réelles. Cela permet aux services de progresser sur un *PoC* tout en préparant et sécurisant les données réelles pour l'industrialisation.

2.4. Définir des intentions pour éclairer les décisions des agents *GenAI* et conserver le contrôle

Certains types d'agents d'IA, notamment ceux relevant des catégories délibératives, de classification ou autres, prennent et exécutent des décisions. Ils nécessitent donc une connaissance approfondie du contexte dans lequel ils agissent afin d'éviter des prises de décisions hors contexte. Il est donc nécessaire que, sous l'égide des organes de gouvernance, un travail soit entrepris afin d'accompagner l'adoption et le déploiement à grande échelle des agents *GenAI*. Ce travail devrait inclure les actions suivantes :

- **Transcrire la raison d'être et la mission de l'organisation** en règles contextuelles utiles aux agents de prise de décisions, et traduire les contraintes réglementaires en doctrines claires.
- **Hiérarchiser les objectifs** : productivité, agilité, acceptabilité sociale et sociétale, impact environnemental.
- **Identifier de nouvelles métriques**, notamment en matière de bien-être et d'acceptabilité, et déterminer les seuils d'alerte au sein d'un dialogue social élargi, ouvrant à une certaine transparence.

HUB FRANCE IA

3. Responsabilités de la gouvernance des agents GenAI

3. Responsabilités de la gouvernance des agents *GenAI*

3.1. Évaluation et audit des systèmes d'IA

L'une des responsabilités fondamentales de la gouvernance des agents *GenAI* est l'évaluation et l'audit de ces derniers. Cela implique la mise en place de processus d'audit réguliers pour examiner les résultats produits par ces agents, notamment le niveau d'hallucination et l'acceptabilité. Ces audits permettent d'identifier les problèmes potentiels et d'assurer la conformité avec les besoins métiers, les normes éthiques et réglementaires.

Les exigences de l'*AI Act* se chevauchent souvent avec celles du Règlement Général sur la Protection des Données (RGPD)[16]. Les entreprises doivent intégrer des principes de « *Privacy by Design* » dans leurs agents d'IA, effectuer des évaluations d'impact pour les agents *GenAI* à haut risque et maintenir une documentation claire de protection des données.

3.2. Évaluation et gestion des risques

Les entreprises doivent effectuer des évaluations de risques pour classer leurs agents d'IA en fonction de leur niveau de risque (inacceptable, élevé, limité, minimal) mais aussi en fonction de l'impact potentiel. Cela permet de prioriser les efforts de conformité et de s'assurer que les agents *GenAI* à haut impact soient conformes aux exigences de l'*AI Act* mais aussi au bon fonctionnement de l'entreprise de manière durable.

La gestion des risques associés à l'utilisation des agents *GenAI* est une autre responsabilité cruciale[17]. Les organisations doivent identifier les risques, tels que les biais algorithmiques, la sécurité des données et les impacts notamment sur la vie privée. Cela

[16] Règlement (UE) 2016/679 du Parlement européen et du Conseil du 27 avril 2016 relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données, et abrogeant la directive 95/46/CE (règlement général sur la protection des données) (Texte présentant de l'intérêt pour l'EEE) https://eur-lex.europa.eu/legal-content/FR/TXT/?uri=celex%3A32016R0679

[17] Owasp. Agentic AI - Threats and Mitigations - OWASP Top 10 for LLM Apps & Gen AI. 2025. https://genai.owasp.org/resource/agentic-ai-threats-and-mitigations/

nécessite l'élaboration de stratégies pour atténuer ces risques, y compris des évaluations d'impact régulières et des mesures de sécurité renforcées.

3.3. Formation et sensibilisation

Une autre responsabilité importante est la formation et la sensibilisation des employés aux pratiques éthiques et responsables liées à l'utilisation de l'IA. Les organisations doivent s'assurer que leurs équipes comprennent les implications éthiques de l'IA et sont formées pour utiliser ces technologies de manière responsable.

3.4. Engagement des parties prenantes

La gouvernance des agents *GenAI* doit également inclure l'engagement des parties prenantes, notamment les employés, les clients et la société civile. Cela implique de consulter ces groupes lors de l'élaboration de politiques et de pratiques de gouvernance, afin de s'assurer que les préoccupations et les attentes de tous soient prises en compte.

HUB FRANCE IA

4. Acteurs et mesures de performance de la gouvernance des agents GenAI

NOTICE

GOUVERNANCE DES AGENTS EXPERTS D'IA GENERATIVE

Juillet2025

4. Acteurs et mesures de performance de la gouvernance des agents *GenAI*

4.1. Acteurs de la gouvernance

Les principaux acteurs impliqués dans la gouvernance des agents *GenAI* sont les équipes de direction, les responsables de la conformité, les développeurs et concepteurs d'IA, ainsi que des représentants des parties prenantes externes. Les équipes de direction sont responsables de la définition des stratégies et des politiques de gouvernance, tandis que les développeurs doivent intégrer des considérations éthiques dès la phase de conception. L'IA agentique représente un changement organisationnel plus important encore que l'arrivée de l'agilité par rapport au cycle en V dans le monde du développement. L'agilité en informatique était encore optionnelle, notamment dans certains secteurs très réglementés. Or l'IA agentique, la pression concurrentielle et l'incertitude géopolitique imposent un changement de posture radical des gouvernances très hiérarchisées afin d'acquérir davantage d'agilité, de fluidité et même de collaboration entre les différentes instances de gouvernance existantes.

Le secteur bancaire, utilisateur majeur d'IA classique, est indéniablement précurseur dans l'organisation de sa gouvernance IA, notamment pour se conformer à des réglementations telles que la SR 11-7[18]. Cette réglementation est cruciale car elle garantit la précision, la fiabilité et la gestion rigoureuse des modèles quantitatifs utilisés par les institutions financières pour la prise de décision. Le non-respect de SR 11-7 peut entraîner des sanctions réglementaires, des dommages à la réputation et des pertes financières liées à l'utilisation de modèles défectueux.

Afin de se conformer à la réglementation SR 11-7 sur la gestion des risques liés aux modèles, les banques ont mis en place une structure organisationnelle fondée sur trois lignes de défense :

- **Première ligne de défense**. Elle regroupe les développeurs et propriétaires de modèles ainsi que les unités opérationnelles. Ils sont responsables du développement, de l'implémentation et de la gestion quotidienne des modèles

[18] Board of Governors of the Federal Reserve System, Washington, D.C. 20551, Division of Banking Supervision and Regulation SR 11-7 April 4, 2011
https://www.federalreserve.gov/supervisionreg/srletters/sr1107.htm

- **Deuxième ligne de défense**. Elle est constituée des équipes de validation indépendante et de gouvernance des modèles. Ces fonctions établissent les normes, supervisent et coordonnent la gestion des risques sans être impliquées dans le développement.
- **Troisième ligne de défense**. Représentée par l'audit interne, elle évalue de façon indépendante l'efficacité du cadre de gestion des risques et vérifie la conformité des deux premières lignes aux politiques et exigences réglementaires.

Cette structure tripartite assure une séparation claire des responsabilités, garantit la transparence et l'efficacité dans la gestion des risques de modèles, et permet une supervision adéquate par la direction générale et le conseil d'administration, conformément aux exigences de la SR 11-7.

La Société Générale a par exemple déployé ce principe de trois lignes de défense pour gérer les risques des cas d'usage basés sur des IA.

4.2. Indicateurs de conformité et indicateurs nouveaux

Les organisations doivent établir des indicateurs de conformité pour évaluer le respect des réglementations et des normes éthiques. Cela inclut la vérification de la conformité avec des lois telles que le RGPD et d'autres réglementations pertinentes. Des indicateurs nouveaux, notamment de bien-être, d'impact social et sociétal, doivent être introduits, mais avec la contrainte imposée par l'*AI Act* de ne pas utiliser les émotions à l'encontre des employés[19].

4.3. Évaluations de l'impact

Les évaluations de l'impact mesurent l'influence des systèmes d'intelligence artificielle (SIA) sur les processus organisationnels et les parties prenantes. Elles peuvent inclure des analyses sur la manière dont les décisions prises par l'IA influencent les résultats commerciaux, la satisfaction des clients et l'engagement des employés[20]. Des indicateurs d'impact environnementaux seront également intégrés à l'évaluation régulière de la bonne utilisation de certaines techniques d'IA, notamment celles moins frugales.

---

[19] Hub France IA. Fiche *AI Act*– Définitions clés du Hub France IA : https://www.hub-franceia.fr/wp-content/uploads/2024/09/Definitions-AI-Act_Analyse_Hub_France_IA.pdf

[20] *Op cit 13.*

4.4. Mesures de la qualité des données

La qualité des données est un aspect crucial de la gouvernance des agents *GenAI*. Les organisations doivent surveiller des indicateurs tels que l'exactitude, l'exhaustivité et la pertinence des données utilisées pour entraîner les modèles d'IA[21]. Il ne s'agit plus de la qualité des données traditionnellement requise pour la comptabilité. L'IA peut tolérer des critères de qualité de données moins stricts, mais cela dépend évidemment des techniques d'IA et de l'impact des défauts de qualité.

4.5. Suivi des biais algorithmiques

Le suivi des biais algorithmiques est essentiel pour garantir l'équité et l'objectivité des SIA. Les organisations doivent mettre en place des mécanismes pour détecter et atténuer les biais dans les modèles d'IA, en utilisant des métriques spécifiques pour évaluer l'équité des résultats produits par ces systèmes[22] et leurs impacts, notamment en fonction des « intentions » et objectifs poursuivis.

4.6. Satisfaction des parties prenantes

La satisfaction des parties prenantes, y compris des employés et des clients, est un indicateur important de la performance de la gouvernance des agents *GenAI*. Des enquêtes et des retours d'expérience peuvent être utilisés pour évaluer la perception des utilisateurs concernant la transparence, le partage des responsabilités et l'efficacité des systèmes d'IA[23].

---

[21] Reuben Binns. Fairness in Machine Learning: Lessons from Political Philosophy. Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency PMLR, 149-158. 2018. https://proceedings.mlr.press/v81/binns18a/binns18a.pdf

[22] *Op cit 10.*

[23] *Op cit 13.*

Conclusion

On constate que certaines organisations intègrent à leur gouvernance générale une gouvernance spécifique à l'intelligence artificielle (IA) générative et aux agents *GenAI*. La gouvernance de ces agents représente un domaine complexe qui exige une attention particulière de la part des organisations. Une adoption à grande échelle de l'IA agentique ne peut se faire sans une gouvernance rigoureuse, ne se limitant pas aux aspects de retour sur investissement, et ne tenant pas compte des risques résiduels à gérer dans le temps. Ces risques incluent une part intangible liée à l'acceptation des changements organisationnels induits, souvent sous-estimée. Cependant, en définissant des objectifs clairs, des responsabilités bien établies, des métriques complémentaires, non seulement financières, mais aussi relatives à l'acceptabilité, à l'éthique et à l'impact environnemental, en impliquant les acteurs concernés et en mesurant tous les aspects de la performance de manière rigoureuse, les organisations peuvent évoluer efficacement dans le contexte complexe de l'IA agentique tout en assurant une utilisation agile, éthique et responsable de ces technologies.

Remerciements

Le Hub France IA remercie l'ensemble des participants au groupe de travail IAG, et tout particulièrement les contributeurs de ce livrable.

Les pilotes

- **Georges Acar**, Inquizyt
- **Alberto Tépox**, Hub France IA

Les contributeurs

- **Benjamin Bosch**, Société Generale
- **Bertrand Lafforge**, Konverso
- **Henry Peyret**, Wassati
- **Hugo David**, Mindflow

Relecture

- **Clodéric Mars**, CM Labs Simulations
- **Emmanuel Adam**, INSA Hauts-de-France
- **Florent Carlier**, Université de Le Mans
- **Matthieu Boussard**, Craft.ai
- **Maxime Morge**, Université Claude Bernard Lyon 1
- **Nicolas Sabouret**, Université Paris-Saclay, CNRS, LISN
- **Valérie Renault**, Université de Le Mans
- **Wassila Ouerdane**, Centrale Supélec
- **Zahia Guessoum**, Université de Reims

Validation

- **Françoise Soulié-Fogelman**, Hub France IA
- **Caroline Chopinaud**, Hub France IA
- **Pierre Monget**, Hub France IA

La touche finale

- **Mélanie Arnould**, Hub France IA

NOTICE

GOUVERNANCE DES AGENTS EXPERTS D'IA GENERATIVE

Juillet 2025

HUB FRANCE IA