Publications du Fellowship en politiques IA

Opérationnalisation de la Loi sur l'intelligence artificielle de l'UE pour les systèmes d'IA multiagents

Maartje Nugteren

Collaborations et remerciements :

Autrice : Maartje Nugteren est chercheuse en politiques d'IA à Mila, l'Institut québécois d'intelligence artificielle, et gestionnaire de programme technique au sein de l'équipe d'IA responsable de Microsoft AI. La recherche qu'elle a menée dans le cadre de son fellowship porte sur la manière dont la Loi sur l'intelligence artificielle de l'UE peut être mise en œuvre pour les systèmes d'IA multiagent.

Collaborateurs : Damiano Fornasiere et Matt MacDermott ont supervisé les recherches et ont fourni des conseils, des commentaires et des révisions inestimables tout au long du fellowship.

Remerciements : Marta Bieńkiewicz (Cooperative AI Foundation), Ben Bariach, Kailey Zengo, et Philipp Schoenegger ont fourni de précieux commentaires et révisions. L'autrice remercie Isadora Hellegren, Laëtitia Vu et Dan Munro pour leurs commentaires, leur soutien éditorial et la gestion du Fellowship Mila en politiques de l'IA, ainsi que Julia Smakman (Ada Lovelace Institute) pour avoir coorganisé l'atelier d'experts sur l'IA multiagent et la Loi sur l'intelligence artificielle de l'UE qui a servi de base à cette note politique, organisé par la Délégation générale du Québec à Bruxelles.

Mila

Introduction

À mesure que le déploiement des systèmes multiagents prend de l'ampleur, il devient urgent de comprendre les défis que posent les systèmes multiagents au cadre existant de la Loi sur l'intelligence artificielle de l'Union européenne. Cette note politique propose des mesures concrètes pour rendre la Loi opérationnelle dans ces contextes.

Les systèmes d'IA multiagents représentent un changement de paradigme fondamental dans le déploiement de l'IA. Alors que les systèmes d'IA traditionnels fonctionnent comme des outils discrets réagissant aux entrées de la personne qui les utilise, les environnements multiagents génèrent des comportements à partir d'interactions entre des agents autonomes. Les principaux protocoles d'interopérabilité, comme Agent2Agent[1] de Google et Model Context Protocol[2] d'Anthropic, permettent une coordination à grande échelle entre les fournisseurs. Ces évolutions créent un écosystème dans lequel les systèmes d'IA fonctionnent, de plus en plus, non pas de manière isolée, mais en tant que participants à des réseaux complexes d'agents en interaction, ce qui soulève de nouvelles questions sur la manière dont les risques émergent et sur la responsabilité de leur gestion.

Les agents d'IA sont construits en équipant les modèles d'IA à usage général (IAUG) de capacités d'appel d'outils, d'accès à des systèmes externes et d'instructions permettant un comportement autonome et orienté vers un objectif. Cela crée une chaîne de valeur dans laquelle différents protagonistes peuvent être responsables du modèle sous-jacent, de l'application d'agent construite sur ce modèle et de son fonctionnement dans le monde réel. Par exemple, l'agent « Claude Code » d'Anthropic combine un modèle fondateur (Claude) avec une couche d'application orientée vers le développeur ou la développeuse et des configurations de déploiement contrôlées par la personne utilisatrice, chacune étant potentiellement gouvernée par différentes personnes ayant des responsabilités distinctes. Si les risques liés aux systèmes multiagents résultent des interactions au moment du déploiement, les facteurs qui déterminent ces risques — les capacités des modèles, le choix de conception des

[1] Fondation Linux. (2025). Linux Foundation Launches the Agent2Agent Protocol Project to Enable Secure, Intelligent Communication Between AI Agents. https://www.linuxfoundation.org/press/linux-foundation-launches-the-agent2agent-protocol-project-to-enable-secure-intelligent-communication-between-ai-agents.
[2] Anthropic. (2024). Introducing the Model Context Protocol. https://www.anthropic.com/news/model-context-protocol.

systèmes et les configurations de déploiement — sont répartis sur l'ensemble de cette chaîne de valeur.

Toutes les configurations multiagents ne posent pas les mêmes problèmes de gouvernance. De nombreuses interfaces d'IA actuelles impliquent déjà une orchestration interne, déléguant le traitement des documents à des modèles plus petits, ou coordonnant plusieurs agents sur des tâches complexes sous le contrôle d'un fournisseur unifié. Cette note politique se concentre principalement sur les interactions entre les fournisseurs ou les organisations, dans lesquelles aucune des parties prenantes ne dispose d'une visibilité totale.

Les systèmes d'IA multiagents promettent des avantages considérables : ils permettent de résoudre des problèmes sophistiqués grâce à une intelligence répartie, d'automatiser une coordination complexe à grande échelle et de faciliter la collaboration des agents spécialisés pour relever des défis qu'un seul système ne pourrait gérer, notamment en étendant le calcul de l'inférence à de multiples agents en parallèle. Toutefois, les systèmes multiagents créent également des modes de défaillance qui diffèrent fondamentalement des risques liés à un système unique. Les problèmes de coordination se produisent lorsque les agents ne parviennent pas à coopérer malgré des objectifs alignés, ce qui entraîne des défaillances séquentielles ou des conflits de ressources. Un conflit survient lorsque des agents ayant des objectifs concurrents se nuisent mutuellement ou nuisent à de tierces parties. La collusion apparaît lorsque des agents se coordonnent d'une manière qui porte préjudice aux consommateurs et consommatrices, aux concurrents et concurrentes ou à l'intérêt public[3].

[3] Hammond, L., Chan, A., Clifton, J. et coll. (2025). Multi-agent risks from advanced AI. Cooperative AI Foundation et Université de Toronto. arXiv:2502.14143.

Les risques liés à l'interaction des agents ne sont pas théoriques. Les spécialistes ont combiné deux modèles indépendamment sûrs — Claude 3 Opus et Llama 2 70B — et ont constaté que la paire générait des logiciels vulnérables dans 43 % des cas, contre moins de 3 % pour chaque modèle séparément. Aucun des deux modèles n'a violé les spécifications de sécurité ; la défaillance n'est due qu'à leur interaction[4]. Ces risques d'interaction ont des précédents dans d'autres domaines de l'automatisation algorithmique. Le krach éclair de 2010 a montré comment les interactions entre des systèmes de négociation autonomes pouvaient produire une instabilité en cascade : un seul ordre de vente a déclenché une dynamique d'autorenforcement entre plusieurs algorithmes, effaçant temporairement près de 1 000 milliards de dollars de valeur de marché en quelques minutes, sans qu'aucun système individuel n'ait dysfonctionné. Ces résultats montrent que des risques importants peuvent provenir non seulement d'agents individuels, mais aussi d'agents agissant ensemble ou à contre-courant[5] [6].

Ces défis sont, de plus en plus, reconnus sur le plan international. Des recherches récentes montrent que l'IA agentive pose des problèmes de responsabilité qui dépassent les cadres traditionnels de gestion des risques par une seule organisation[7]. Dans le cadre de gouvernance modèle de Singapour pour l'IA agentive, il est noté que, lorsque plusieurs agents interagissent, des risques peuvent survenir au niveau du système, les erreurs d'un agent se répercutant rapidement sur les autres[8]. Le Rapport international sur la sécurité de l'IA de 2026 confirme que les systèmes multiagents introduisent de nouveaux modes de défaillance, avec des erreurs qui peuvent se propager et s'amplifier à travers les interactions entre les agents[9].

La Loi sur l'intelligence artificielle de l'UE, qui constitue le cadre réglementaire le plus complet en matière d'IA actuellement mis en œuvre, présente à la fois un test critique et une occasion d'établir des approches de gouvernance dont d'autres juridictions pourraient s'inspirer.

---

[4] Jones, E., Dragan, A. et Steinhardt, J. (2024). Adversaries Can Misuse Combinations of Safe Models. arXiv:2406.14595.
[5] Dignum, V. et Dignum, F. (2025). Agentifying Agentic AI. Prépublication arXiv. arXiv:2511.17332.
[6] U.S. Securities and Exchange Commission et Commodity Futures Trading Commission. (2010). Findings Regarding the Market Events of May 6, 2010. https://www.sec.gov/news/studies/2010/marketevents-report.pdf.
[7] Center for Long-Term Cybersecurity, UC Berkeley. (2026). Agentic AI Risk-Management Standards Profile.
[8] Infocomm Media Development Authority of Singapore (2026). Model AI Governance Framework for Agentic AI Systems.
[99] International AI Safety Report. (2026). https://www.internationalaisafetyreport.org.

Nous ne partons pas de zéro

Le code de bonnes pratiques de l'IAUG reconnaît déjà les risques liés aux agents multiples :

- ➔ Il cite la collusion et la mauvaise coordination comme étant des propensions du modèle, qu'il faut évaluer.
- ➔ Il exige des mesures d'atténuation en matière de sécurité, y compris des techniques pour des écosystèmes sûrs d'agents d'IA.
- ➔ Il énumère les capacités pertinentes, notamment l'utilisation d'outils, les interactions avec d'autres systèmes d'IA et l'accès aux infrastructures critiques[10].

Le défi consiste à traduire ces reconnaissances en normes d'essai concrètes, en exigences en matière de documentation et en mécanismes de coordination — en particulier pour les interactions qui dépassent les frontières des fournisseurs.

La question du champ d'application

La Loi sur l'intelligence artificielle de l'UE régit les systèmes d'IA en fonction de l'usage auquel ils sont destinés et de la classification des risques qu'ils peuvent entraîner. Les systèmes à haut risque relevant du chapitre III sont soumis à des obligations exhaustives : gestion des risques, robustesse, transparence, surveillance humaine, surveillance après la mise en marché et notification des incidents. Les modèles d'IAUG présentant un risque systémique en vertu du chapitre V doivent faire l'objet d'une évaluation et d'une atténuation supplémentaires. Pour les agents concernés, la question est de savoir comment opérationnaliser les obligations existantes dans des contextes multiagents.

De nombreux agents d'IA ne relèvent cependant pas de ces catégories, ce qui signifie que les exigences en matière de gouvernance qui s'y appliquent sont minimes[11]. Pourtant, des agents restreints spécifiques à une tâche, sont intégrés dans

[10] General-Purpose AI Code of Practice. (2025).
[11] Loi sur l'intelligence artificielle de l'UE, art. 6, annexe III.

l'infrastructure numérique — dans les applications grand public, les services en ligne et les systèmes d'entreprises — où ils interagissent régulièrement avec d'autres agents ou services. Lorsque ces agents à risque limité interagissent entre eux ou influencent les résultats en amont des décisions à risque élevé, ils peuvent créer des risques collectifs qu'aucun fournisseur individuel n'est tenu d'évaluer ou d'atténuer.

Les agents d'IA doivent-ils être classés par défaut dans la catégorie des agents à haut risque ?

Comme l'affirme The Future Society, les nouveaux mécanismes de risque introduits par les agents — l'action autonome, l'utilisation d'outils, la coordination multiagent — peuvent justifier une mise à jour de l'annexe III. La Commission est habilitée à adopter des actes délégués ajoutant des catégories à haut risque (article 7), et plusieurs critères semblent pertinents : l'autonomie, le préjudice potentiel et l'absence d'autres mécanismes de recours[12]. Yousefi et coll. proposent un mécanisme de « hausse du risque » : si un système atteint un degré élevé d'autonomie agentive, il devrait automatiquement déclencher une classification à haut risque, que son domaine soit ou non répertorié à l'annexe III[13]. Ce principe sous-jacent — que les seuils d'autonomie peuvent déclencher de manière indépendante la classification des risques, peu importe le domaine — mérite d'être sérieusement pris en considération lorsque la Commission exerce le pouvoir qui lui est délégué en vertu de l'article 7, en particulier dans le contexte du réexamen de l'annexe III prévue à l'article 112. Toutefois, l'opérationnalisation nécessiterait des nuances. Deux dimensions sont particulièrement importantes : le degré de proactivité, c'est-à-dire le fait que les agents n'agissent qu'en réponse à des invites explicites ou qu'ils peuvent entreprendre des actions à distance temporelle de l'instruction initiale, et la portée et la sensibilité des outils accessibles, puisqu'un agent restreint interrogeant un moteur de recherche diffère catégoriquement d'un agent ayant accès à des systèmes financiers ou à des moyens de communication. La classification des risques devrait tenir compte de l'interaction entre ces facteurs plutôt que de traiter l'autonomie comme une caractéristique binaire.

[12] Oueslati, A. et Staes-Polet, R. (2025). Ahead of the Curve: Governing AI Agents under the EU AI Act. The Future Society.
[13] Yousefi, Y., Billi, M. et Rotolo, A. (2025). Agentic AI: An EU AI Act Paradigm Shift? Alma AI, Université de Bologne.

En mai 2026, le projet de lignes directrices de la Commission sur la classification des systèmes d'IA à haut risque a commencé à répondre à ces questions, confirmant que lorsque des composants modulaires se combinent pour influencer une décision à haut risque, l'ensemble de la configuration est évalué comme un système d'IA unique, et les modules individuels ne peuvent pas échapper à la classification. Cette approche s'étend expressément aux systèmes agentiques dont les actions liées entre elles servent un objectif à haut risque. La mesure anti-contournement est bienvenue, mais elle fonctionne dans le cadre des objectifs existants de l'annexe III : elle capture les configurations qui servent conjointement un objectif unique à haut risque, et non les risques collectifs qui émergent lorsque des agents sous contrôle séparé interagissent à travers les frontières des fournisseurs, où aucune configuration individuelle ne sert un tel objectif. Une approche fondée sur le seuil d'autonomie en vertu de l'article 7 reste donc complémentaire plutôt que redondante[14].

Tous les risques liés aux systèmes multiagents ne relèvent pas du champ d'application de la Loi sur l'intelligence artificielle, et ils ne devraient pas non plus en relever. Certains risques, comme les scénarios de collusion ressemblant à un comportement de cartellisation, peuvent être traités de manière plus appropriée par le droit de la concurrence. D'autres peuvent nécessiter une réglementation sectorielle dans les domaines de l'énergie, des soins de santé ou des services financiers. Une gouvernance efficace nécessite une coordination entre les instruments : la Loi sur l'intelligence artificielle, le RGPD, le droit de la concurrence et les cadres sectoriels jouent des rôles complémentaires.

Cette note politique se concentre sur les risques pouvant être traités par les mécanismes de mise en œuvre de la Loi sur l'intelligence artificielle, tout en reconnaissant l'existence d'un panorama multi-instruments plus large. Les sections suivantes examinent les défis que représente la mise en œuvre des systèmes à risque élevé fondés sur des modèles d'IAUG, pour lesquels les obligations prévues par la Loi sont les plus claires, mais dont l'opérationnalisation dans des contextes multiagents reste sous-développée.

[14] Commission européenne. (2026). Projet de lignes directrices de la Commission sur la classification des systèmes d'IA à haut risque (UE) 2024/1689. Publié le 19 mai 2026 ; consultation ouverte jusqu'au 23 juin 2026.

Les défis de la mise en œuvre

La Loi sur l'intelligence artificielle de l'UE prévoit des obligations pour l'ensemble de la chaîne de valeur : les entités qui fournissent des modèles d'IAUG, celles qui fournissent les systèmes à haut risque et celles qui les déploient ont chacune des responsabilités distinctes. Ces dispositions ont été conçues principalement pour des systèmes et des modèles individuels.

La Loi reconnaît que les systèmes interagissent : le paragraphe 72(2) exige explicitement des fournisseurs qu'ils analysent les interactions. Néanmoins, le cadre part du principe que chaque fournisseur est en mesure d'évaluer correctement les risques dans le cadre de son propre champ de contrôle. Dans les environnements multiagents, cette hypothèse ne tient pas. Comme le dit le Gradient Institute, « une collection d'agents sûrs ne garantit pas une collection sûre d'agents »[15]. Trois défis se présentent.

Les trois défis de la mise en œuvre

Risques émergents	Visibilité fragmentée	Rapport cloisonné
Deux agents sûrs peuvent produire ensemble des effets néfastes.	Aucun fournisseur ne voit comment son système interagit avec les autres.	Les dynamiques d'interaction n'apparaissent pas dans un seul et même rapport.

Ces défis se manifestent tout au long du cycle de vie du déploiement, des tests et de la documentation avant la mise en marché à la réaction aux incidents, en passant par la surveillance après la mise en marché.

[15] Reid, T., Schroeder, C. et Thompson, N. (2025). Risk Analysis Techniques for Governed LLM-based Multi-Agent Systems. Gradient Institute.

Avant la mise en marché

Les tests

Les fournisseurs de systèmes à risque élevé doivent cerner et analyser les risques raisonnablement prévisibles, tandis que les fournisseurs d'IAUG doivent évaluer les risques systémiques (articles 9 et 55). Le code de bonnes pratiques détaille ces obligations, y compris les tests de propension à la collusion et à la mauvaise coordination[16]. Cependant, les tests de propension évaluent les tendances intrinsèques des modèles individuels : ils ne peuvent pas détecter des dynamiques qui n'existent que lorsque les modèles interagissent.

Cela crée une incertitude sur le plan pratique. Si le déploiement multiagent entre fournisseurs est prévisible — et compte tenu des protocoles d'interopérabilité, il l'est de plus en plus —, que faut-il faire pour respecter l'article 9 ? Les fournisseurs ne peuvent pas tester tous les partenaires d'interaction possibles. Pourtant, prétendre que les risques d'interaction sont imprévisibles devient de moins en moins défendable au fur et à mesure que le déploiement de systèmes multiagents devient pratique courante.

La difficulté est encore aggravée par l'apprentissage continu. Contrairement aux logiciels traditionnels qui restent statiques après leur déploiement, les systèmes agentiques peuvent adapter leur comportement en fonction des interactions — que ce soit par des mécanismes de mémoire explicite ou par l'apprentissage au cours de l'exécution d'une tâche prolongée[17]. La Loi sur l'intelligence artificielle de l'UE définit la « modification substantielle » comme étant un changement qui affecte la conformité et qui n'a pas été prévu ou planifié dans l'évaluation initiale de la conformité (paragraphe 3[23]), mais cette définition suppose que les modifications sont des événements discrets plutôt qu'une adaptation continue. Des orientations sont nécessaires pour déterminer quand l'apprentissage autonome constitue une modification substantielle nécessitant un réexamen.

La documentation

La documentation fait face à des défis similaires. Les articles 53 et 54 exigent des fournisseurs d'IAUG qu'ils documentent leurs capacités et leurs limites ; l'article 13 exige des fournisseurs de systèmes à haut risque qu'ils fournissent aux entreprises de

[16] International AI Safety Report. (2026).
[17] Dignum et Dignum. (2025).

déploiement des informations sur les caractéristiques des systèmes. Pourtant, les fournisseurs ne peuvent pas documenter des comportements qui résultent uniquement d'interactions avec des systèmes externes qu'ils ne contrôlent pas. Même lorsque les fournisseurs documentent ce qu'ils peuvent observer, ces limitations documentées peuvent décrire des environnements contrôlés, mais sans saisir le comportement émergeant des interactions avec les systèmes externes. L'évaluation de modèles isolés est donc insuffisante pour comprendre comment ces modèles se comporteront dans des environnements interactifs[18].

Ce qui est nécessaire

Les orientations devraient clarifier la manière dont les obligations existantes s'appliquent aux contextes multiagents. Lorsque le déploiement de systèmes multiagents est prévisible, la gestion des risques doit tenir compte des scénarios d'interaction : il ne s'agit pas de tester tous les partenaires possibles, mais d'examiner systématiquement le comportement sous la pression de la concurrence, les instructions contradictoires et la coordination avec des systèmes qui poursuivent des objectifs différents. La responsabilité devrait être répartie en fonction de la visibilité : les fournisseurs de modèles documentent les propriétés d'interaction connues et indiquent explicitement les cas où des essais multiagents n'ont pas été effectués ; les fournisseurs de systèmes et les entreprises de déploiement, qui voient les contextes de déploiement réels, sont les premiers responsables des tests dans leurs environnements spécifiques.

Les normes techniques en vertu de l'article 40 devraient inclure des protocoles pour une évaluation axée sur l'interaction et des tests en bac à sable multifournisseurs. Compte tenu de leur complexité, les normes devraient être élaborées par l'entremise de projets pilotes concrets avant d'être codifiées. Cela permettra d'apprendre par la pratique.

Après la mise en marché

La transparence pour les personnes utilisatrices

Lorsqu'elles interagissent avec un système d'IA, les personnes utilisatrices doivent en être informées (article 50). Le projet de lignes directrices de la Commission sur les

[18] Dignum et Dignum (2025) ; Center for Long-Term Cybersecurity (2026).

obligations de transparence en vertu de l'article 50, publié en mai 2026, confirme que les systèmes agentiques entrent dans le champ d'application et que l'interaction est désormais entendue comme couvrant les échanges d'actions, et pas seulement de contenu. Lorsqu'un fournisseur ne peut pas déterminer de façon fiable si un agent interagira avec une personne physique, l'agent doit divulguer sa nature artificielle chaque fois qu'une telle interaction est probable. Cela concerne le cas de l'agent à l'humain, mais pas le cas multiagent, dans lequel plusieurs agents façonnent ensemble un seul résultat[19]. Dans les contextes multiagents les interactions conséquentes peuvent se produire entre des systèmes d'IA et être invisibles pour les personnes concernées jusqu'à ce que les dommages se matérialisent. La personne utilisatrice voit une interface, sans savoir que l'agent a délégué des requêtes à des spécialistes, consulté des systèmes externes ou fait modifier les résultats par un traitement en aval. Le processus de prise de décision est réparti entre les agents d'une manière que la divulgation d'un seul système n'exprime pas[20].

Les spécifications des entreprises de déploiement

Les entreprises de déploiement sont confrontées à un déficit de spécifications. Elles doivent utiliser les systèmes à haut risque conformément aux instructions du fournisseur et surveiller les risques (article 26), mais la Loi ne prévoit pas de cadre permettant aux entreprises de déploiement de formuler des exigences multiagents avant l'approvisionnement. Une entreprise de déploiement qui intègre un agent d'IA à des systèmes automatisés existants ne peut pas facilement spécifier les exigences de coordination, les attentes en matière de journalisation ou les conditions limites, et ne dispose d'aucun moyen normalisé pour vérifier ces propriétés après le déploiement. Cela dépend en partie des obligations du fournisseur : les entreprises de déploiement ne peuvent pas spécifier ou vérifier le comportement multiagent si les fournisseurs ne sont pas, en premier lieu, tenus de documenter les caractéristiques relatives à l'interaction.

Le contrôle

L'obligation d'analyser les interactions (paragraphe 72[2]) représente un progrès, mais elle s'inscrit dans le cadre d'un fournisseur unique. Chaque fournisseur analyse les

[19] Commission européenne. (2026). Projet de lignes directrices sur la mise en œuvre des obligations de transparence pour certains systèmes d'IA au titre de l'article 50 de la législation sur l'IA Publié le 8 mai 2026 ; consultation ouverte jusqu'au 3 juin 2026.
[20] Loi sur l'intelligence artificielle de l'UE, art. 50.

interactions auxquelles son système est confronté ; il n'existe aucun mécanisme permettant de partager les résultats lorsque la dynamique néfaste s'étend à plusieurs systèmes. Il en résulte une visibilité fragmentée : chaque partie prenante ne voit que sa portion, tandis que les schémas à l'échelle du système ne sont pas détectés[21].

Cela crée un problème de traçabilité. Comme les agents agissent de manière autonome, par exemple en effectuant des achats, en envoyant des communications et en exécutant du code, il devient essentiel de pouvoir déterminer quels systèmes ont contribué à l'obtention d'un résultat. Pourtant, les mécanismes actuels ne permettent pas aux fournisseurs de reconstituer les agents impliqués dans une interaction donnée, ce qui limite l'efficacité de la surveillance post-commercialisation[22].

Ce qui est nécessaire

Les obligations en vertu de l'article 50 devraient s'étendre aux contextes multiagents : lorsque plusieurs agents contribuent à un résultat, les fournisseurs devraient divulguer la nature distribuée de la prise de décision. Plus fondamentalement, la transparence multiagent nécessite des mécanismes d'identification des agents. Les agents d'IA devraient porter des identifiants vérifiables permettant à d'autres systèmes et à une enquête d'établir quels agents ont participé à une interaction[23]. Au-delà de l'identification, des systèmes de certification pourraient fournir une vérification par une tierce partie des propriétés de l'agent, comme les outils accessibles, le degré d'autonomie et les pratiques de traitement des données, ce qui permettrait aux autres parties prenantes de prendre des décisions éclairées en matière d'interaction. Des approches techniques émergent, notamment les identités numériques sécurisées, les identificateurs d'instance unique et les journaux d'activité vérifiables[24]. Celles-ci n'en sont encore qu'à leurs débuts, ce qui souligne la nécessité de mener des projets pilotes avant de procéder à la codification. La journalisation structurée des interactions avec les agents devrait devenir pratique courante et offrir une documentation suffisante pour reconstituer les séquences d'interaction en cas d'incident, avec une profondeur et une conservation qui sont proportionnelles au risque.

[21] Reid et coll. (2025).
[22] Chan, A., Wei, K., Huang, S. et coll. (2025). Infrastructure for AI agents d'IA Agents. arXiv:2501.10114.
[23] Chan et coll. (2025).
[24] Chan et coll. (2025).

La réaction aux incidents

L'attribution

Les fournisseurs de systèmes à haut risque et de modèles d'IAUG présentant un risque systémique doivent signaler les incidents graves (article 73 et paragraphe 55[1]). Ces dispositions supposent que les incidents peuvent être attribués à des systèmes et à des fournisseurs précis. Cependant, les incidents multiagents peuvent impliquer des défaillances séquentielles entre les fournisseurs, dans lesquelles aucun système individuel ne fonctionne mal, mais où le comportement collectif cause des dommages.

Des recherches récentes confirment que l'attribution des défaillances dans les systèmes multiagents est fondamentalement difficile. Zhang et coll. (2025) ont constaté que même l'IA de pointe n'atteint qu'une précision de 53,5 % dans l'identification de l'agent à l'origine de l'échec d'une tâche, l'attribution au niveau de l'étape n'atteignant que 14,2 %. Les spécialistes en chair et en os ont eu besoin de plus de 30 heures et de plusieurs cycles de consensus pour annoter les journaux de défaillance, 15 à 30 % des cas étant marqués comme incertains[25]. Si l'attribution est aussi difficile avec un accès complet aux journaux et à l'analyse de spécialistes, les fournisseurs ne peuvent raisonnablement pas diagnostiquer eux-mêmes les défaillances multiagents de manière isolée.

Les rapports fragmentés

L'architecture de la Loi en matière de rapports fragmente encore davantage le tableau. Sur le plan vertical, les fournisseurs d'IAUG relèvent du bureau de l'IA, tandis que les fournisseurs de systèmes à haut risque relèvent des autorités nationales de surveillance du marché. Lorsqu'un incident implique à la fois la propension d'un modèle et des choix de déploiement au niveau du système, les informations pertinentes sont réparties dans des canaux distincts. Horizontalement, lorsque les systèmes de plusieurs fournisseurs interagissent pour produire un préjudice, chaque fournisseur ne signale que ce qu'il peut observer à partir de ses propres registres. Par conséquent, lorsqu'un incident multiagent se produit, aucune autorité ne reçoit une image complète de la dynamique d'interaction qui a causé le dommage.

[25] Zhang, Z. et coll. (2025). Which Agent Causes Task Failures and When? Automatic Failure Attribution in Multi-Agent LLM Systems. arXiv:2505.00212.

La surveillance à la vitesse de l'éclair

Le défi est d'autant plus grand que la vitesse est élevée. L'exécution autonome peut aller plus vite que la surveillance et la réactivité. La cascade rapide du krach éclair illustre la rapidité avec laquelle les interactions peuvent s'emballer[26]. Même lorsqu'une surveillance humaine existe, les approches existantes reposent largement sur la vérification rétrospective et l'inspection des journaux, dont l'efficacité est limitée pour les systèmes fonctionnant à grande vitesse[27]. Lorsque plusieurs agents interagissent de manière autonome, ce défi s'accroît. La vitesse et le volume des échanges entre agents peuvent dépasser toute capacité humaine d'examen digne de ce nom. En outre, comme les agents communiquent de plus en plus par l'intermédiaire de représentations apprises plutôt que de textes lisibles par l'être humain, ce que l'on appelle parfois le « neuralese », la langue de l'interaction entre les agents peut elle-même devenir impénétrable pour les êtres humains, ce qui aggrave le problème de la surveillance.

Les signalements par de tierces parties

Les cadres actuels ne prévoient pas non plus de voies pour l'établissement de rapports par de tierces parties. Les personnes utilisatrices, la société civile ou les universitaires peuvent observer des défaillances multiagents que les fournisseurs ont peu de chances de détecter parce qu'il s'agit de schémas plus visibles de l'extérieur du système, comme des comportements corrélés entre les fournisseurs, des impacts cumulatifs sur des populations précises ou des effets sur l'ensemble du marché qu'aucun journal d'un seul fournisseur ne pourrait révéler. Le projet de lignes directrices en vertu de l'article 73 ne prévoit aucun mécanisme permettant à ces observations d'entrer dans le processus réglementaire[28].

Ce qui est nécessaire

Les protocoles d'enquête interfournisseurs sont essentiels. Lorsqu'un comportement préjudiciable résulte d'interactions entre des systèmes, aucun fournisseur ne peut à lui seul procéder à une analyse adéquate des causes profondes. Le bureau de l'IA devrait élaborer des cadres d'enquête coordonnée, dont les spécificités seraient développées de manière itérative par l'entremise de projets pilotes. Les mécanismes devraient

[26] U.S. SEC et CFTC (2010).
[27] Yousefi et coll. (2025).
[28] Fernández Ashman, N., Anwar, U. et Bieńkiewicz, M. (2026, 13 janvier). EU Regulations Are Not Ready for Multi-Agent AI Incidents. Tech Policy Press.

également permettre à la société civile, aux universitaires et aux personnes utilisatrices de soumettre des notifications étayées d'incidents multiagents présumés, ces schémas étant souvent plus visibles de l'extérieur du système qu'à partir des journaux d'un seul fournisseur.

Conclusions

Comme le souligne le Rapport international sur la sécurité de l'IA, le rythme des progrès de l'IA pose un « dilemme des preuves » : attendre des preuves plus solides des risques pourrait laisser la société sans préparation[29]. Les écosystèmes multiagents risquent de s'enraciner et d'être plus difficiles à gérer une fois qu'ils auront été déployés à grande échelle. Cette analyse arrive à un moment critique : les normes techniques sont en cours de rédaction et, en mai 2026, la Commission a publié son premier projet de lignes directrices sur la classification des risques élevés et sur la transparence en vertu de l'article 50. C'est la première fois que la Commission engage ces dispositions à la lumière de l'IA agentive. L'accord politique sur le train de mesures « omnibus numérique » sur l'IA a reporté l'application des règles à haut risque, prolongeant plutôt que fermant la fenêtre pour obtenir une mise en œuvre multiagent correcte. Les décisions prises aujourd'hui façonneront la gouvernance d'une technologie qui est déjà en train de prendre de l'ampleur.

Nous ne partons pas de zéro et nous n'avons pas besoin d'attendre des preuves parfaites. Le cadre fondé sur les risques de la Loi sur l'intelligence artificielle, la reconnaissance par le code de bonnes pratiques des risques liés aux systèmes multiagents et l'émergence d'un consensus international constituent une base solide. Le défi et l'occasion à saisir consistent à traduire ces éléments en mécanismes de mise en œuvre pratique par l'intermédiaire d'expérimentations, de projets pilotes et de l'apprentissage par la pratique.

Si nous voulons que les agents se coordonnent en toute sécurité, nous devons nous coordonner nous-mêmes. Le bureau de l'IA, les organismes de réglementation sectoriels, les autorités de la concurrence et les partenaires internationaux ont tous un rôle à jouer. La gouvernance multiagent, comme l'IA multiagent elle-même, nécessite une coordination qui dépasse les frontières.

[29] IMDA Singapour (2026).

Clause de non-responsabilité

Les opinions exprimées dans cet article sont strictement celles des auteurs et ne représentent ni ne reflètent nécessairement les politiques ou positions officielles de Mila, de ses affiliés, de ses administrateurs ou de ses bailleurs de fonds. Les auteurs assument l'entière responsabilité de l'exactitude et de l'intégrité de ce travail.

Des perspectives à concrétiser

Les recommandations suivantes traduisent l'analyse ci-dessus en mécanismes de mise en œuvre concrets, réalisables par les canaux existants : orientations en vertu de l'article 96, demandes de normalisation en vertu de l'article 40 et renforcement des capacités du bureau de l'IA.

Domaine	No	Recommandation	Responsabilité
Portée	**R1**	Examiner si l'atteinte de certains seuils d'autonomie ou de capacité par les agents d'IA justifie de les ajouter aux catégories à haut risque de l'annexe III, compte tenu des nouveaux mécanismes de risque (action autonome, utilisation d'outils, coordination multiagent) et des critères énoncés au paragraphe 7(2). Les directives de classification de mai 2026 portent sur les configurations modulaires et agentiques servant un seul objectif à haut risque ; une loi déléguée permettrait en outre de couvrir les risques liés à l'autonomie et à l'interaction, que l'interprétation seule ne peut pas cerner.	Commission européenne
Avant la mise en marché	**R2**	Exiger des tests axés sur l'interaction pour les systèmes à haut risque et les modèles d'IAUG dans lesquels le déploiement multiagent est prévisible, y compris des protocoles d'évaluation pour la coordination émergente, des scénarios d'interaction antagoniste et des essais en bac à sable multifournisseurs.	Commission européenne

	R3	Soutenir l'élaboration de normes de communication sécurisées et interexploitables entre agents par l'intermédiaire du CEN/CENELEC ou d'organismes multipartites, couvrant les identifiants d'agents vérifiables, les protocoles d'authentification et les exigences en matière de journalisation entre fournisseurs.	CEN/ CENELEC/ ETSI
	R4	Exiger des fournisseurs qu'ils documentent les caractéristiques relatives aux interactions : comportements sous pression concurrentielle et coopérative, protocoles de communication utilisés, limites des espaces d'action des agents et mises en garde explicites lorsque des tests multiagents n'ont pas été réalisés.	Commission européenne
Après la mise en marché	**R5**	Préciser que les obligations de transparence prévues à l'article 50 s'étendent aux contextes multiagents, en exigeant la divulgation lorsque plusieurs agents contribuent à un résultat.	Commission européenne
	R6	Élaborer des orientations en matière d'approvisionnement et des modèles de spécifications permettant aux entreprises de déploiement de formuler des exigences en matière de comportement multiagent — identification des agents, mesures de protection quant à la coordination, capacités de journalisation.	Bureau de l'IA + ENISA
Réaction aux incidents	**R7**	Élaborer des orientations sur la coordination des enquêtes en cas de comportement préjudiciable résultant de l'interaction des systèmes, y compris (a) les critères permettant de déterminer quand la coordination entre les fournisseurs est justifiée ; (b) les attentes minimales en matière de partage d'informations entre les fournisseurs ; (c) la clarification des rôles de coordination du bureau de l'IA lorsque les incidents relèvent de plusieurs juridictions ; et (d) explorer, par l'entremise de projets pilotes volontaires, ce à quoi pourrait ressembler en pratique la divulgation après enquête. Les normes doivent être élaborées de manière itérative avant d'être codifiées.	Commission européenne/ bureau de l'IA

	R8	Mettre en place des mécanismes permettant à la société civile, aux chercheurs et aux personnes utilisatrices de soumettre des notifications étayées en cas d'incidents multiagents présumés.	Bureau de l'IA (avec mise en œuvre par les États membres)
Institutionnel	**R9**	Mettre en place une expertise multiagent au sein du bureau de l'IA, des mécanismes de coordination avec les organismes de réglementation sectoriels et les autorités de la concurrence, ainsi qu'une coopération internationale en matière de normes.	Bureau de l'IA/ Commission européenne + États membres