CNIL.
COMMISSION NATIONALE
INFORMATIQUE & LIBERTÉS

Conseil
de l'IA
et du Numérique

Note

IA agentique et protection des données personnelles : équation à inconnues multiples pour les utilisateurs

juillet 2026

Résumé opérationnel

Le passage à l'agentique marque un changement d'échelle dans les usages d'intelligence artificielle (IA), lequel amplifie les risques de cybersécurité – qui ont fait l'objet d'une note précédente du CIANum – ainsi que les risques pour les données personnelles des utilisateurs, sujet de cette note. En effet, en dépassant le simple traitement de requêtes pour atteindre une autonomie d'action croissante, les systèmes agentiques questionnent la mise en œuvre du cadre légal de la protection des données et diluent la maîtrise de l'utilisateur sur la circulation de ses données personnelles.

Caractérisées par leur capacité à réaliser de multiples tâches successives et à interconnecter plusieurs agents au système orchestrateur, les IA agentiques constituent une rupture dans le traitement et la circulation des données personnelles. Elles mettent ainsi en tension les principes fondamentaux du Règlement général sur la protection des données personnelles (RGPD). Pour autant, ces principes essentiels restent absolument essentiels à la protection des données personnelles, sujet crucial à l'ère de l'intelligence artificielle afin d'éviter des dérives sociétales majeures.

Cette note s'attache à examiner l'application du RGPD à l'IA agentique, tandis que l'application complète du Règlement sur l'IA (RIA), prévue en 2027, ne prévoit pas de règles spécifiques pour les agents IA.

Le fonctionnement et l'amélioration permanente des agents IA nécessitent une connaissance fine, continue et évolutive de l'utilisateur. Grâce à des mécanismes de mémoire persistante, l'IA agentique crée des profils utilisateurs très précis qui, couplés à une autonomie décisionnelle accrue, font peser le risque d'une perte de maîtrise de l'utilisateur sur ses données personnelles.

De plus, les flux, parfois opaques et souvent entre de multiples services, complexifient l'attribution des responsabilités des différents acteurs de la chaîne de valeur. Le manque de transparence de cette architecture accentue les risques liés aux décisions automatisées, en particulier dans des secteurs sensibles ou régulés (financiers ou judiciaires, par exemple).

Enfin, cette note identifie des voies de conciliation possibles pour garantir l'effectivité du cadre du juridique et le déploiement des systèmes d'IA agentique, par la mise place de mécanismes techniques de contrôle et de supervision. Sur le plan juridique, cela nécessitera des règles de transparence renforcées sur l'action agentique ou l'obligation d'une validation humaine pour les décisions les plus critiques *(human in the loop*). Sur le plan technique, il sera important d'intégrer des mesures dès la conception des systèmes, telles que le cloisonnement de la mémoire des agents pour éviter l'accumulation de données, le déploiement dans

des environnements isolés (*sandboxing*), ou encore l'installation d'un bouton d'arrêt d'urgence (*kill switch button*) à la main de l'utilisateur.

Le développement de systèmes d'IA respectueux des utilisateurs est un enjeu majeur qui doit préoccuper l'ensemble de l'écosystème afin de tirer le meilleur bénéfice de ces nouvelles technologies tout en nous préservant collectivement des nombreux risques susceptibles de se matérialiser.

Table des matières

Introduction

Le développement de l'IA agentique constitue une évolution importante dans les interactions entre les utilisateurs et les systèmes d'IA actuellement répandus. On désigne par IA agentique un ensemble de programmes informatiques reposant sur des modèles d'IA générative et capables de prendre des décisions de manière autonome, d'orchestrer des actions complexes et d'interagir avec des services tiers, avec ou sans validation humaine[1] .

L'architecture d'un système d'IA agentique repose généralement sur l'interaction entre plusieurs composants :

(i) Un **agent « orchestrateur »**, au cœur duquel le modèle d'IA générative permet à l'utilisateur d'interagir en langage naturel avec le système. Ce modèle constitue ainsi l'interface principale entre l'utilisateur et le système.

(ii) Des **agents « spécialisés »**, coordonnés par l'agent orchestrateur, qui peuvent accomplir des tâches complexes nécessitant des compétences spécifiques (génération de code, traitement de texte, paiements en ligne, *etc.*).

(iii) L'ensemble de ces agents peut ensuite être connecté à des services externes comme des applications, des bases de données ou des outils de recherche *web* qu'ils peuvent mobiliser pour traiter les requêtes.

Les interactions entre l'agent orchestrateur, les agents spécialisés et les services tiers **reposent sur des protocoles d'échange standardisés**[2], qui établissent un langage de communication commun entre les modèles d'IA et les applications tierces. Ces protocoles sont indispensables à l'interopérabilité des différents systèmes et permettent aux agents d'accéder aux ressources externes nécessaires à la réalisation des actions attendues.

*

Pour assurer la continuité des interactions avec l'utilisateur et exécuter les tâches qui leur sont confiées, les systèmes d'IA agentique mettent en œuvre deux mécanismes spécifiques de gestion des données :

- Le « **processus** », c'est-à-dire le traitement de requêtes, repose sur un « **contexte** » mobilisé par l'agent, lui permettant de conserver l'historique des échanges avec l'utilisateur, les instructions reçues (directement par l'utilisateur ou par un autre agent), ainsi que l'ensemble des interactions avec les autres agents et services impliqués. Le contexte est supprimé à la fin du processus.

---

[1] Pour plus d'informations voir la première note de la série IA agentique du Conseil de l'IA et du numérique. *Les intelligences artificielles à l'heure de la vague agentique : de quoi parle-t-on ?*. Février 2026.

[2] Les protocoles MCP, développé par Anthropic, et ACP, développé par OpenAI, sont des exemples de protocoles très utilisés :
https://modelcontextprotocol.io/docs/getting-started/intro
https://www.agenticcommerce.dev/docs

- En complément, le « **stockage mémoire** » permet de conserver des informations réutilisables d'un processus à l'autre, notamment des données relatives à l'utilisateur. Cette mémoire peut être alimentée ou modifiée par l'utilisateur, ou enrichie automatiquement à partir des interactions avec le système d'IA agentique. Elle est persistante et indépendante des processus d'exécution.

Chaque agent (orchestrateur ou spécialisé) dispose de son contexte propre et de sa propre mémoire, regroupant les informations utiles pour l'exécution de ses tâches. Dans certains cas, il peut également exister une **mémoire partagée et commune à plusieurs agents**.

*

Les systèmes d'IA agentique sont susceptibles de traiter des volumes importants de données personnelles, y compris des données sensibles au sens du RGPD[3], ainsi que des informations confidentielles. Or, le cadre posé par le RGPD, conçu à une époque où les traitements étaient principalement déterminés et mis en œuvre par des acteurs humains dans la poursuite de finalités préalablement définies, est désormais appelé à s'appliquer à des environnements dans lesquels une partie des opérations de traitement peut être initiée, coordonnée ou exécutée de manière autonome par des systèmes d'IA agentique.

Si les **IA agentiques peuvent conserver des données personnelles de manière prolongée**[4], elles sont également en mesure de **créer des profils utilisateur hyper-personnalisés.** Par extension, en agissant de manière autonome, sans intervention directe de l'utilisateur, l'IA agentique peut être amenée à prendre des décisions ou à accomplir des actes susceptibles de produire des effets juridiques à l'égard de celui-ci.

De telles capacités soulèvent des interrogations fondamentales quant à la délégation de la prise de décision par des systèmes autonomes, à la responsabilité des actes qu'ils accomplissent, ainsi qu'au maintien d'un contrôle effectif par l'utilisateur ou par des tiers sur leur fonctionnement.

Dans cette perspective, cette note examine comment les principes fondamentaux du RGPD sont mis en tension dans le cadre de l'IA agentique (I), capable d'autonomie croissante et d'hyperpersonnalisation, au risque d'une perte de maîtrise des utilisateurs de leurs données personnelles (II). Cette autonomie engendre en outre une opacité des flux décisionnels, laquelle remet en cause les régimes de responsabilité et de sécurité (III). Enfin, cette note s'attache à identifier les voies de conciliation entre le développement de ces technologies et les exigences de protection des données (IV).

---

[3] Article 9-1 du RGPD : Sont considérées comme données sensibles les *« données à caractère personnel qui révèlent l'origine raciale ou ethnique, les opinions politiques, les convictions religieuses ou philosophiques ou l'appartenance syndicale, ainsi que le traitement des données génétiques, des données biométriques aux fins d'identifier une personne physique de manière unique, des données concernant la santé ou des données concernant la vie sexuelle ou l'orientation sexuelle d'une personne physique »* .

[4] BHOSALE Robini *et al.* The dark side of autonomous intelligence: a survey on data leakage and privacy failures in agentic AI. *Frontiers in Computer Science*. 2026. vol. 8, p. 4.

I. De l'IA générative à l'IA agentique : vers l'amplification des risques pour les données personnelles

(a) L'opacité des flux de données augmente les risques pour l'utilisateur

Les IA agentiques partagent un socle commun de vulnérabilités avec les IA génératives en matière de protection des données personnelles. Cependant, leur capacité à accéder à de multiples sources d'information, à conserver un contexte étendu, à mobiliser de nombreux outils externes et à exécuter des actions de manière autonome et avec une grande liberté entraîne une **multiplication des flux de données**, tant au sein du système d'IA agentique (échange de données entre agents) qu'en dehors de celui-ci (actions réalisées auprès de services externes, telles que l'envoi d'un courrier électronique ou une transaction financière pour l'acquisition d'un bien, *etc.*).

Du point de vue de l'utilisateur, il n'est pas toujours évident d'appréhender l'ampleur des données traitées, ni d'en comprendre précisément les finalités. Par ailleurs, il est particulièrement difficile de percevoir la nécessité de chacune des interactions, d'autant plus que les échanges entre les différents agents peuvent être nombreux, notamment en cas d'échecs répétés. Ces flux peuvent alors devenir importants et conduire à des transferts de données massifs entre les différents services.

(b) Les effets de l'IA agentique : une mise en tension des principes du RGPD

Les systèmes d'IA agentique peuvent complexifier l'identification, la traçabilité et la maîtrise des traitements de données personnelles. Sans remettre en cause la pertinence du cadre posé par le RGPD, **cette technologie soulève des difficultés particulières quant à la mise en œuvre concrète de plusieurs de ses principes fondamentaux :**

Le principe de finalité (article 5.1.b) exige que les données personnelles soient collectées pour des objectifs déterminés, explicites et légitimes[5]. Or les systèmes d'IA agentique sont conçus pour réaliser un grand nombre de tâches dans des domaines variés, ce qui peut rendre plus difficile l'identification du périmètre exact des traitements effectués. L'enjeu consiste à garantir que les données traitées par l'agent ne soient utilisées que pour l'exécution des tâches qui lui sont confiées et qu'elles ne fassent pas l'objet de réutilisations incompatibles avec les finalités initialement définies.

Le principe de licéité (article 6) impose que tout traitement repose sur un fondement juridique valable, à choisir parmi l'une des six bases légales prévues par le Règlement européen[6]. L'autonomie des systèmes d'IA agentique peut rendre plus complexe le lien entre les nouvelles opérations réalisées et la base juridique initialement choisie.

---

[5] Article 5, paragraphe 1, point b), du RGPD.

[6] Parmi les six bases légales prévues par le RGPD : le consentement, le contrat, l'obligation légale, la mission d'intérêt public, l'intérêt légitime et la sauvegarde des intérêts vitaux.

Le principe de minimisation (article 5.1.c) exige que les données traitées soient adéquates, pertinentes et limitées à ce qui est nécessaire au regard des objectifs poursuivis. Pourtant, les IA agentiques peuvent mobiliser de grandes quantités de données (e-mails, historique de navigation, fichiers), les partager entre agents et les conserver en mémoire afin de répondre aux besoins de l'utilisateur, voire de les anticiper. La nécessité de traiter l'ensemble de ces données n'est pas toujours évidente à démontrer.

Le principe d'exactitude (article 5.1.d) prévoit que les données doivent être exactes et, si nécessaire, mises à jour. Or, la nature probabiliste des modèles d'IA générative implique un risque d'erreur. Ces hallucinations peuvent se propager dans les systèmes sans que l'utilisateur ne dispose de mécanisme d'alerte avec des conséquences potentiellement importantes pour l'utilisateur.

Le principe de transparence (article 5.1.a) exige que les personnes concernées soient informées de manière claire et accessible des données traitées les concernant. Par ailleurs, les sorties produites par l'IA générative sont difficiles à expliquer en raison de l'architecture des modèles sous-jacents et de leur fonctionnement probabiliste[7].

Le principe de limitation de la durée de conservation des données (article 5.1.e) impose que les données personnelles ne soient conservées que pendant une durée limitée et proportionnée aux objectifs poursuivis. Cette exigence peut toutefois entrer en tension avec certaines pratiques telles que la conservation de jeux de données destinés à l'entraînement ou à l'amélioration des modèles d'IA générative, ainsi qu'avec la capacité de ces derniers à mémoriser certaines informations issues des données utilisées lors de leur développement[8]. Qui plus est, les instances mémoire diffuses et multiples des systèmes d'IA agentiques complexifient le contrôle de ces durées.

*

Enfin, le RGPD garantit aux **personnes concernées[9] l'exercice de plusieurs droits** sur leurs données personnelles, notamment les droits d'accès, de rectification, d'effacement, d'opposition et de limitation du traitement (article 15 à 22 du RGPD). **L'exercice effectif de ces droits soulève toutefois des difficultés particulières dans le contexte des systèmes d'IA fondés sur des modèles d'IA générative.** La CNIL a déjà pu relever que l'identification et la suppression des données personnelles au sein des données d'entraînement des modèles d'IA générative peut s'avérer complexe[10]. Ces difficultés sont encore accentuées dans les systèmes d'IA agentique, en raison de la circulation des données entre plusieurs agents, espaces de mémoire et services tiers. Concrètement, la personne concernée peut avoir du mal à savoir quel agent a collecté l'information, où elle a été conservée, si elle a été transmise à un autre composant du système, et auprès de quel responsable elle doit exercer ses droits. La rectification d'une

[7] LINC. « *L'explicabilité de l'IA à l'ère de l'apprentissage profond* ». 8 juillet 2025.

[8] VEALE Michael, BINNS Reuben, EDWARDS Lilian. "*Algorithms that Remember: Model Inversion Attacks and Data Protection Law*". Cornell University. 2018.

[9] Une personne concernée est une personne au sujet de laquelle des données à caractère personnel sont traitées.

[10] CNIL. « *IA : Respecter et faciliter l'exercice des droits des personnes concernées* ». 7 février 2025.

donnée erronée ou son effacement peuvent ainsi devenir difficiles à mettre en œuvre de manière complète et traçable.

II. Autonomie et hyperpersonnalisation de l'IA agentique : quelles incidences sur la maîtrise des données personnelles ?

Les capacités d'hyperpersonnalisation et d'exploitation de grandes quantités de données des IA agentiques sont de nature à renforcer les risques de délégation de la décision et, par conséquent, d'occasionner une perte de maîtrise de l'utilisateur.

(a) La constitution d'une connaissance approfondie de l'utilisateur par les systèmes agentiques

L'une des principales caractéristiques de l'IA agentique réside dans sa capacité à construire une connaissance approfondie et évolutive de l'utilisateur. L'autonomie de ces systèmes ne repose pas uniquement sur les capacités des modèles d'IA générative, mais également sur l'accumulation progressive de données permettant d'anticiper les besoins de l'utilisateur et d'agir en son nom. **Cette logique conduit à un changement d'échelle dans le traitement des données personnelles.** Celles-ci ne sont plus seulement mobilisées pour répondre à une requête ponctuelle mais sont continuellement conservées, enrichies et réutilisées au sein du système. **Ce fonctionnement est à la fois vecteur d'une efficacité inédite, tant les profils utilisateurs sont précis et le système autonome, et catalyseur de risques relatifs à la maîtrise des utilisateurs sur leurs données personnelles.** Cette accumulation de données résulte principalement de deux mécanismes complémentaires : **la mémoire persistante des agents**, d'une part, et l'exploitation de sources de données externes, d'autre part.

(1) La mémoire des agents : un outil de personnalisation fondé sur la conservation des données personnelles

Les mécanismes de mémoire constituent **une brique centrale du fonctionnement des systèmes d'IA agentique**, en permettant notamment aux agents d'adapter leurs réponses et d'améliorer l'exécution des tâches qui leur sont confiées. Toutefois et comme vu en introduction, dans les architectures impliquant plusieurs agents, la multiplication des espaces de mémoire peut conduire à une dispersion des données entre différents environnements de stockage. **Cette organisation décentralisée peut accroître l'opacité des traitements, en rendant plus complexe la détermination des données effectivement conservées, des opérations susceptibles de les affecter (ajout, modification, suppression) ainsi que de leur durée de conservation.**

> Par exemple, si un utilisateur supprime un fichier sensible de son ordinateur dans un dossier partagé avec son IA agentique, il peut être difficile pour lui de s'assurer que les données personnelles ont bien été effacées de l'ensemble de la mémoire

de l'IA agentique. Si ces données sont conservées par le système d'IA agentique, elles sont susceptibles de continuer à être utilisées lors d'interactions futures.

Par ailleurs, la coexistence de plusieurs mémoires au sein d'un même système soulève des difficultés de synchronisation. Des informations divergentes, contradictoires ou obsolètes peuvent se propager entre les agents. Or, ces flux de données sont susceptibles d'avoir des conséquences sur l'exactitude des données traitées. En effet, la qualité des données constitue un enjeu déterminant dès lors que le système s'appuie sur celles-ci pour exécuter une tâche ou prendre une décision.

(2) L'enrichissement du profil de l'utilisateur par l'exploitation de sources de données externes

Afin de répondre aux requêtes, le système d'IA agentique peut compléter les informations conservées dans sa mémoire par des données issues de services connectés (services tiers, *web etc.*). Ces données permettent d'enrichir le **contexte** (cf. introduction) des agents IA. En pratique, le recours aux services connectés est souvent initié par le système d'IA agentique lui-même et n'est pas nécessairement explicité dans la requête de l'utilisateur, ce qui renouvelle les enjeux liés au respect du RGPD, déjà évoqués, en lien avec l'opacité des flux de données.

À titre d'exemple, lorsqu'un utilisateur donne à une IA agentique l'accès à son environnement de travail (messagerie électronique, agenda, espaces documentaires ou outils collaboratifs), celle-ci peut consulter l'ensemble des informations disponibles afin de répondre à une requête donnée. Toutefois, certaines de ces informations peuvent ne pas être strictement nécessaires à l'exécution de la tâche demandée. **Les données ainsi collectées peuvent ensuite être intégrées à différents espaces de mémoire ou transmises à des services externes mobilisés dans le cadre du traitement de la requête**.

La capacité des agents à croiser des données issues de sources multiples permet une collecte évolutive et cumulative de données personnelles, au fil d'interactions hyper-personnalisées. En d'autres termes, le système d'IA agentique peut, à partir des informations dont il dispose, déduire toujours plus d'informations sur la vie privée de l'utilisateur, qui seront susceptibles d'être mobilisées (parfois implicitement) pour exécuter des nouvelles requêtes.

(b) De l'assistance à la délégation décisionnelle

L'autonomie de ces systèmes est donc susceptible de créer un déséquilibre dans la relation entre l'utilisateur et son IA agentique, traduisant une forme implicite de délégation du pouvoir décisionnel de l'utilisateur vers le système d'IA agentique. Ce constat est partagé notamment par la présidente de Signal, Meredith Whittaker, qui décrit les promesses de l'IA agentique par

la formulation imagée suivante : *« put your brain in a jar and let AI handle life »*[11] (ou *« mettez votre cerveau dans un bocal et laissez l'IA gérer votre vie »*). En effet, à partir des sources de données auxquelles il peut accéder et aux connaissances qu'il accumule sur l'utilisateur, **un système d'IA agentique ne se limite plus à un rôle d'assistant à la décision. Il peut exécuter de manière autonome l'ensemble des étapes nécessaires à la réalisation d'une requête, réduisant ainsi progressivement le besoin d'intervention humaine**[12].

Pour autant, l'autonomie accrue des IA agentiques et la difficulté d'interrompre des processus fonctionnant en continu peuvent engendrer des analyses erronées des requêtes ainsi que des risques d'hallucinations en cascade susceptibles de se propager au sein du système et d'impacter la mémoire des agents et les actions prises.

> Une employée d'une grande entreprise du numérique a rapporté dans un article récent la suppression, par son IA agentique, de nombreux e-mails professionnels et les difficultés pour interrompre le processus à distance[13]. Si cet exemple ne remet pas en cause le principe même de l'automatisation, il illustre les enjeux liés à la fiabilité des décisions prises par des systèmes autonomes : une mauvaise interprétation de la requête initiale ou une erreur dans l'analyse peut conduire l'agent à effectuer des opérations inadaptées sur des volumes importants de données. Dans ce contexte, la capacité à identifier rapidement l'origine d'une action, à la corriger et à limiter ses effets devient un élément essentiel de la maîtrise du système.

L'autonomie de ces systèmes est susceptible de conduire à la prise de décisions relevant, dans certaines hypothèses, du **régime des décisions individuelles automatisées** prévu par l'article 22 du RGPD. Celui-ci reconnaît à toute personne le droit de ne pas faire l'objet d'une décision fondée exclusivement sur un traitement automatisé, y compris le profilage, produisant des effets juridiques la concernant ou l'affectant de manière significative de façon similaire. Une telle décision ne peut être licite qu'en présence de l'une des exceptions prévues au paragraphe 2 du même article (exécution d'un contrat, autorisation par le droit de l'Union ou d'un État membre, consentement explicite).

Au regard de la multiplicité des actions et des agents IA impliqués dans l'exécution d'une tâche, **l'appréciation du degré réel de supervision dans le processus décisionnel** demeure une question délicate. L'existence seule d'une intervention humaine en sortie dans la décision proposée par un système d'IA ne suffit pas nécessairement à écarter la qualification de décision fondée exclusivement sur un traitement automatisé au sens de l'article 22 du RGPD. Comme l'a jugé la CJUE dans son arrêt sur le score SCHUFA[14], l'intervention humaine doit être

---

[11] TechCrunch. "Agentic AI is like putting your brain in a jar". 7 mars 2025.
[12] DEFFAINS Bruno. « L'IA agentique : du changement d'échelle technologique au questionnement du modèle juridique classique ». *Dalloz Actualités* IP/IT. Novembre 2025.
[13] Business Insider. "Meta alignement director shares OpenClaw email deletion nightmare".
[14] CJUE. Affaire SCHUFA C-634/21. 7 décembre 2023.

réelle, effective et exercer une influence sur la décision finale ; une validation purement formelle ou automatique est insuffisante.

III. L'opacité des flux décisionnels dans l'IA agentique : flou autour de la responsabilité des acteurs et sécurité des données

L'autonomie décisionnelle des systèmes d'IA agentique ne soulève pas uniquement des interrogations quant à la licéité des traitements de données personnelles ; elle remet également en cause les mécanismes traditionnels d'imputation des responsabilités et de maîtrise des risques. En multipliant les interactions entre modèles, agents et services tiers, ces systèmes rendent plus difficile l'identification des acteurs responsables d'un traitement ou d'un dommage, tout en élargissant les surfaces d'exposition aux failles de sécurité.

*

Au regard du RGPD, le déploiement de systèmes d'IA agentique ne remet pas en cause le principe selon lequel un responsable du traitement doit pouvoir être identifié et demeurer en mesure d'assurer le respect des obligations en matière de protection des données personnelles. Toutefois, les caractéristiques propres à ces systèmes, notamment leur capacité à mobiliser plusieurs agents, à interagir avec des services tiers et à prendre de manière autonome certaines décisions, peuvent complexifier l'identification des responsabilités respectives des différents acteurs. Cette complexité est susceptible de soulever des difficultés particulières en matière de gouvernance, de répartition des rôles entre responsables du traitement et sous-traitants, ainsi que de démonstration du respect du principe de responsabilité (ou *accountability*) prévu par le RGPD.

Au-delà du seul régime relatif à la protection des données personnelles, **les décisions des IA agentiques s'inscrivent dans des écosystèmes complexes impliquant une multiplicité d'acteurs** (développeurs, fournisseurs de modèles, intégrateurs, déployeurs, utilisateurs) **et dans lesquelles les chaînes de responsabilité civiles, voire pénales, demeurent floues**. Par exemple, en cas d'erreur (achat erroné, envoi d'un courriel au mauvais destinataire ou divulgation de données confidentielles à un système tiers), il est difficile d'identifier l'origine du dysfonctionnement et, par conséquent, d'établir les responsabilités. **L'identification du régime de responsabilité suppose de pouvoir retracer les différentes étapes du processus décisionnel et de déterminer le rôle respectif des agents et des utilisateurs impliqués.**

Si un système d'IA est dépourvu de personnalité juridique et ne peut, en tant que tel, voir sa responsabilité engagée, il ne saurait pour autant être assimilé à un simple outil passif, en raison de son niveau d'autonomie et de la multiplicité des acteurs intervenant dans sa conception, son développement, son déploiement et son utilisation. Cette difficulté illustre en partie les divergences ayant conduit à l'abandon, en 2025, du projet de directive européenne relative à

la responsabilité en matière d'intelligence artificielle[15]. Aussi, le Règlement sur l'IA (RIA) n'instaure en l'état pas de régime de responsabilité exprès mais impose des obligations de conformité selon une approche fondée sur les risques et selon les acteurs de la chaîne de valeur, ce qui peut permettre d'identifier la responsabilité de chacun. Néanmoins, tout en ne consacrant pas l'IA agentique comme une catégorie spécifique d'IA, les règles du RIA sont applicables à l'IA agentique, ce qui peut permettre d'identifier les rôles de chaque acteur dans la chaîne de valeur. En l'absence de régime spécifique, les règles de droit commun en France demeurent applicables, qu'il s'agisse de la responsabilité délictuelle pour faute (article 1240 du code civil), de la responsabilité du fait des choses (article 1242 du code civil), de la responsabilité contractuelle ou, le cas échéant, du nouveau régime européen applicable aux produits défectueux[16].

*

Au-delà des enjeux juridiques, l'IA agentique présente des risques de sécurité propres à son architecture distribuée : l'interconnexion de plusieurs agents, outils et services accroît la surface d'attaque – une vulnérabilité dans un seul composant peut être exploitée pour affecter l'ensemble du système. La circulation continue des données entre agents, mémoires et services externes crée en outre des points de faiblesse supplémentaires, avec des risques de fuite, de réutilisation non maîtrisée ou de manipulation des informations échangées. Enfin, parce qu'un agent peut agir de manière autonome, une erreur, une tromperie ou une compromission peut le conduire à exécuter une action non souhaitée.

IV. Préserver la maîtrise des données personnelles face aux capacités d'action des IA agentiques : quelles voies de conciliation ?

Au regard des risques précédemment identifiés, **le cadre juridique applicable aux IA agentiques apparaît comme un levier essentiel de maîtrise de ces nouveaux usages**. Si certains instruments du droit européen de la protection des données leur sont déjà applicables, **leurs caractéristiques propres** – autonomie décisionnelle, mémoire persistante, capacité d'interagir avec une pluralité de services et d'agir au nom de l'utilisateur – **justifient une adaptation des modalités de mise en œuvre de ces règles.** Il apparait ainsi particulièrement important, d'une part, de garantir l'effectivité des principes du RGPD au regard des spécificités des IA agentiques et, d'autre part, de renforcer les garanties techniques et la supervision humaine de ces systèmes.

---

[15] Bird & Bird. « Quel(s) régime(s) de responsabilité pour les dommages causés par l'IA ? / CARON Charles-Henri et LAMPE Anne-Sophie ». Tech 360° 3ème édition. Décembre 2025.

[16] Directive (UE) 2024/2853 du 23 octobre 2024, relative à la responsabilité du fait des produits défectueux qui a révisé le cadre issu de la directive de 1985 afin de l'adapter notamment aux technologies émergentes, dont l'intelligence artificielle.

(a) Garantir l'effectivité des principes du RGPD et adapter les recommandations sectorielles

Préciser le cadre juridique applicable à l'IA agentique en particulier.

Il est important de rappeler que les IA agentiques demeurent pleinement soumises au RGPD. Les recommandations élaborées par la CNIL précisent les exigences applicables tant de manière générale aux IA génératives, que dans certains secteurs spécifiques et sensibles, tels que la santé[17] ou l'éducation[18]. Ces recommandations s'adressent, d'une part, aux développeurs de modèles et de systèmes d'IA, notamment dans les fiches pratiques[19] , et, d'autre part, aux déployeurs de ces systèmes[20]. Elles constituent ainsi un premier socle de bonnes pratiques susceptibles de guider également le développement et le déploiement des IA agentiques.

La conformité des systèmes d'IA agentique au droit de la protection des données personnelles pourrait justifier des spécifications quant à l'application de la réglementation à l'IA agentique en particulier. Parmi les travaux récents ou en cours, un projet de recommandations du Conseil de l'Europe[21] a été publié en mai dernier sur les risques liés à la vie privée et à la protection des données associés à l'utilisation de systèmes fondés sur des modèles de langage de grande taille (LLM). Toutefois, ces dernières n'adressent pas tous les défis posés par l'IA agentique et son articulation avec le cadre réglementaire applicable, en particulier le RGPD et le RIA. Des lignes directrices concernant l'articulation de ces deux textes sont actuellement en cours de rédaction par le Comité européen de la protection des données (CEPD) et la Commission européenne[22]. Leur parution est attendue d'ici fin 2026.

S'il est d'ores et déjà acquis que le RIA est applicable aux systèmes d'IA agentique[23], le CEPD pourrait utilement orienter ses futurs travaux en matière d'IA - faisant suite à son avis sur les modèles d'IA [24] et ses récentes lignes directrices sur les pratiques de moissonnage en ligne (*webscraping*)[25] - **vers des recommandations en matière de protection des données adaptées aux développeurs ou déployeurs d'IA agentiques**.

*

[17] CNIL. « *IA et Santé : la HAS et la CNIL lancent une consultation publique sur un projet de guide* ». Mars 2026
[18] CNIL. « *La CNIL publie deux FAQ sur l'utilisation des systèmes d'IA dans les établissements scolaires* ». Juin 2025.
[19] CNIL. Les fiches pratiques IA.
[20] CNIL. Les questions-réponses de la CNIL sur l'utilisation d'un système d'IA générative. Juillet 2024.
[21] Conseil de l'Europe. Draft Guidelines on Privacy and Data Protection in the context of LLM-based systems. Mai 2026.
[22] Commission européenne. « *Soutenir la mise en œuvre de la législation sur l'IA au moyen de lignes directrices claires* ». Décembre 2025.
[23] Commission européenne. « *Comment les agents de l'IA sont-ils traités dans la législation sur l'IA ?* ».
[24] Comité européen sur la protection des données. « *Avis 28/2024 relatif à certains aspects de la protection des données liées au traitement de données à caractère personnel dans le contexte des modèles d'IA* ». Décembre 2024.
[25] Comité européen sur la protection des données. « Guidelines 03/2026 on web scraping in the context of generative AI ». Juillet 2026.

Garantir une meilleure transparence des systèmes d'IA agentique.

L'information délivrée à l'utilisateur doit lui permettre de conserver une compréhension claire et effective des traitements réalisés. Au-delà des obligations générales de transparence prévues par le RGPD, les systèmes agentiques gagneraient à **intégrer des mécanismes de traçabilité permettant de reconstituer l'ensemble du processus décisionnel**. Pour chaque tâche exécutée, l'utilisateur devrait pouvoir identifier les données personnelles mobilisées, les agents intervenus, les services tiers sollicités, les échanges réalisés ainsi que leur chronologie. Une telle documentation contribuerait non seulement à **améliorer la transparence** des décisions prises par le système, mais également à **établir une chaîne de responsabilité** en cas de dysfonctionnement ou d'atteinte aux droits des personnes.

*

Permettre à l'utilisateur de choisir les données auquel le système à accès.

Aussi, cette transparence devrait être complétée par un **contrôle renforcé de l'utilisateur sur les données auxquelles les agents peuvent accéder, en particulier lorsque ces données sont sensibles**. Les principes de minimisation des données et de limitation des finalités imposent en effet que seules les données strictement nécessaires à l'exécution d'une tâche soient traitées.

*

Préciser l'encadrement des décisions automatisées.

Par ailleurs, l'autonomie croissante de ces systèmes invite à rappeler que les garanties prévues par l'article 22 du RGPD demeurent pleinement applicables lorsqu'une décision fondée exclusivement sur un traitement automatisé produit des effets juridiques ou affecte de manière significative la personne concernée. Des recommandations aux spécificités des différents secteurs pourraient utilement préciser les situations dans lesquelles les traitements mis en œuvre par des systèmes d'IA agentique sont susceptibles de relever de ce régime, ainsi que les modalités pratiques de la supervision humaine lorsque celle-ci est requise.

(b) Mettre en place des mécanismes techniques de contrôle et de supervision

Pour mettre en œuvre les garanties juridiques applicables au traitement des données personnelles, la maîtrise des risques liés aux IA agentiques suppose d'agir sur leur architecture technique et leurs capacités d'action. **L'objectif est de limiter les conséquences d'une autonomie excessive en intégrant, dès la conception des systèmes, des mécanismes de contrôle, de limitation et de supervision adaptés aux risques identifiés.**

*

Penser des mesures de détection et de filtrage.

D'abord, des **mesures de détection et de filtrage des usages non prévus ou détournés** doivent être envisagées pour limiter les risques d'usages malveillants. Si aucune méthode n'est infaillible, la combinaison de (i) l'utilisation de modèles entraînés et instruits spécifiquement pour limiter les usages interdits ou illicites et de (ii) la mise en œuvre de mesures de filtrages des requêtes de l'utilisateur permet de réduire ces risques. Des mesures techniques peuvent ainsi être mises en place pour : reconnaître et bloquer les requêtes suspectes, surveiller en

temps réel les actions des agents et alerter en cas de comportement inhabituel, faire des tests de robustesse avant le déploiement et tout au long de celui-ci pour identifier les failles. De telles mesures de filtrage peuvent par ailleurs s'appliquer dès lors qu'un modèle d'IA générative est sollicité à travers un agent (orchestrateur ou spécialisé), et pas uniquement lors de la requête initiale de l'utilisateur qui initie le processus.

*

Cloisonner la mémoire et l'environnement des systèmes.

De plus, **la mémoire des systèmes devrait être cloisonnée** et chaque agent devrait disposer d'une **mémoire dédiée et isolée**, sans accès automatique aux données des autres agents. **Pour limiter les durées de conservation extensives** et assurer la suppression régulière des données, notamment dans les mémoires de chaque agent, des mesures techniques pourraient également être prises. Ces mémoires pourraient être volontairement limitées en taille, les informations contenues expirant automatiquement au bout d'un certain temps, ou étant remplacées régulièrement par des informations plus récentes. Ces mesures peuvent par ailleurs améliorer la tenue à jour des informations si elles sont remplacées régulièrement par des données plus récentes. **Un mécanisme de comparaison des mémoires** entre agents pourrait par ailleurs être implémenté, afin de s'assurer de la cohérence et de la tenue à jour des données personnelles conservées au sujet de l'utilisateur. Enfin, pour réduire les risques de personnalisation excessive de ces systèmes, il peut être recommandé de mette en place un **cloisonnement de la mémoire par traitement afin de limiter le stockage excessif de données**. On peut imaginer la création de sessions dédiées par traitement pour séparer les usages et les données traitées en conséquence. De même, **l'espace de déploiement de l'IA agentique devrait être, lui aussi, cloisonné** (on parle de ***sandboxing***[26]). Un environnement dédié permet de réduire au strict nécessaire les accès aux services et données de l'utilisateur, réduit le périmètre des risques et facilite la surveillance de ces systèmes. Le risque d'exfiltration de données se limiterait alors au périmètre de données susceptibles d'être traitées par l'IA agentique.

*

Conserver l'humain dans la boucle pour les actions critiques.

Il est également possible de **limiter le degré d'autonomie en imposant une intervention humaine pour des conséquences jugées critiques**. À tout le moins, la gouvernance des systèmes autonomes (*AI agents harnessing*) doit constituer une priorité de premier ordre pour les acteurs. Dans cette perspective, l'intégration sécurisée des services tiers (applications, bases de données, outils externes) avec lesquels les systèmes d'IA agentique interagissent représente le principal défi. Des protocoles d'intégration stricts pourraient être mis en œuvre pour limiter les risques liés à leurs interactions. En ce sens, **il peut être utile de classer les actions possibles sur chaque service par niveau de risque** (accès à des données, modification et suppression de donnés, envoi de données hors système et typologie de données traitées : financières, de santé, *etc.*). En fonction du risque évalué, une intervention humaine peut être jugée nécessaire

[26] CNIL. Le *sandboxing « est un mécanisme de sécurité mis en œuvre par un système d'exploitation pour isoler une application exécutée vis-à-vis du cœur du système d'exploitation mais aussi des autres applications exécutées sur le terminal ».*

avant leur exécution. L'utilisateur pourrait par exemple avoir la main sur le choix des actions qui nécessiteraient sa validation avant exécution, pour réduire les risques les plus graves. Par ailleurs, afin de garder le contrôle sur le système d'IA agentique, il peut être envisagé l'intégration d'un *« kill switch »*[27], c'est-à-dire d'une fonctionnalité permettant d'interrompre à tout moment l'exécution d'un processus agentique en cas de comportement inattendu ou de risque identifié.

*

Garantir l'évaluation des IA agentiques.

Enfin, **dans le prolongement des démarches déjà engagées pour l'évaluation indépendante des modèles et des systèmes d'IA**, il serait utile de se pencher davantage **sur l'évaluation de la protection des données personnelles et de la sécurité des IA agentiques**. Sans créer un dispositif entièrement nouveau, une telle évolution permettrait de mieux apprécier leur conformité au RGPD et la robustesse des mesures de protection mises en place. La publication des résultats de ces évaluations pourrait contribuer à renforcer la transparence et la confiance dans ces systèmes. Cette démarche suppose toutefois des moyens dédiés, humains et financiers, et une coordination étroite entre les acteurs concernés.

**

La délicate équation de la protection des données personnelles à l'ère de l'IA agentique est un sujet en évolution permanente et n'est aujourd'hui pas résolue. Des démarches d'interprétation des exigences du RGPD adaptées aux spécificités concrètes de l'IA agentique seraient donc bénéfiques afin de réduire les risques pour les utilisateurs tout en clarifiant le cadre pour les professionnels. La maturité de ces propositions reste bien évidemment à mettre à l'épreuve des innovations constantes dans le secteur.

Il est important de noter que le RGPD n'est pas le seul règlement auquel peuvent être soumis les systèmes d'IA agentique. **Dans une logique distincte, davantage orientée vers la réglementation et la sécurité des produits**, le RIA encadre la mise sur le marché des systèmes d'IA, leur mise en service et leur utilisation selon une approche fondée sur les niveaux de risque. Comme vu *supra*, bien que l'IA agentique ne soit pas explicitement définie par le RIA, ses caractéristiques propres, notamment son autonomie avancée, ses capacités d'action et d'hyperpersonnalisation, pourraient justifier une attention particulière dans son application future. **Certains auteurs**[28] **suggèrent une intégration explicite dans le texte**.

De même, une articulation sera également nécessaire avec les réglementations sectorielles applicables, notamment dans les domaines du travail, de la santé ou de l'éducation, dans lesquels les conséquences d'une délégation accrue à ces systèmes peuvent être particulièrement impactantes pour les individus.

---

[27] Institut Jacques Delors. *Révolution des technologies de l'intelligence artificielle à l'horizon 2035*. Juin 2026.
[28] HACKER Philipp, HOLWEG Matthias. *A Pragmatic Approach to Regulating AI Agents*. Avril 2026.

Au-delà des seules questions de conformité, le développement des systèmes d'IA agentique soulève également des enjeux concurrentiels au sein de cet écosystème. En particulier, les fournisseurs de ces systèmes peuvent, dans certaines conditions, être amenés à réutiliser les données générées ou fournies par leurs utilisateurs afin d'entraîner ou d'améliorer leurs modèles (cf. IV.1). La concentration de ces données entre les mains d'un nombre limité d'acteurs est susceptible de renforcer leur avantage concurrentiel, au « *risque de créer de nouvelles formes de dépendance technologique, voire un « dilemme de souveraineté*[29] »[30], au détriment de la maîtrise des utilisateurs sur leurs données personnelles.

[29] Pour plus d'informations voir : Conseil de l'IA et du numérique. *IA & cybersécurité : anticiper aujourd'hui pour maîtriser demain*. Avril 2026.
[30] BOUVEROT Anne. « Les agents IA, chevaux de Troie pour la sécurité des organisations ? ». *Les Echos*. Juin 2026.