CONSULTATION PUBLIQUE

QUESTIONNAIRE SUR L'APPLICATION DU RGPD AUX MODELES D'IA

SYNTHESE DES CONTRIBUTIONS

Juin 2025

La CNIL a lancé, le 10 juin 2024, une consultation publique sur le développement de systèmes d'IA.

Les contributions ont nourri les travaux du deuxième lot de huit fiches pratiques, dont celle sur l'application du RGPD aux modèles d'IA, pour leur publication définitive sur le site de la CNIL.

CNIL.
COMMISSION NATIONALE
INFORMATIQUE & LIBERTÉS

Questionnaire sur l'application du RGPD aux modèles d'IA Synthèse des réponses

La CNIL a soumis à consultation publique en juin 2024, un questionnaire invitant les fournisseurs et utilisateurs de systèmes d'IA, ainsi que l'ensemble des acteurs concernés, à apporter leurs éclairages sur les conditions dans lesquelles les modèles d'IA peuvent être considérés comme anonymes ou doivent être encadrés par le RGPD.

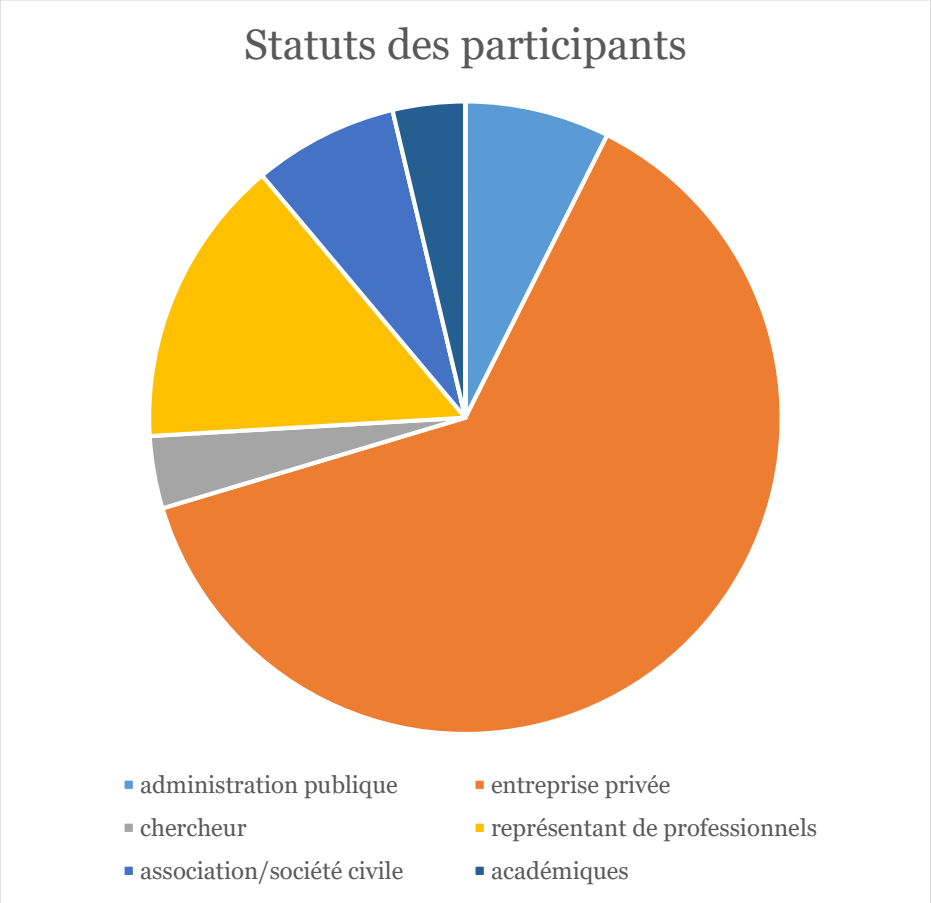
Cette synthèse résume les réponses apportées par les contributeurs concernant :

- leur situation particulière ;
- les risques de réidentification ;
- les techniques permettant d'analyser les risques de régurgitation et d'extraction de données personnelles ;
- l'application du RGPD à un modèle d'IA ayant mémorisé des données personnelles ;
- la responsabilité des acteurs.

1. Sur les répondants aux questionnaires

Le questionnaire sur l'application du RGPD aux modèles d'IA a reçu 27 contributions, provenant de :

- **2 administrations publiques (7%) ;**
- **17 entreprises privées (63%) ;**
- **4 organisations représentantes de professionnels (15%) ;**
- **2 associations représentatives de la société civile (7%) ;**
- **1 chercheur académique en droit (4%) ;**
- **1 regroupement de chercheurs académiques (4%).**

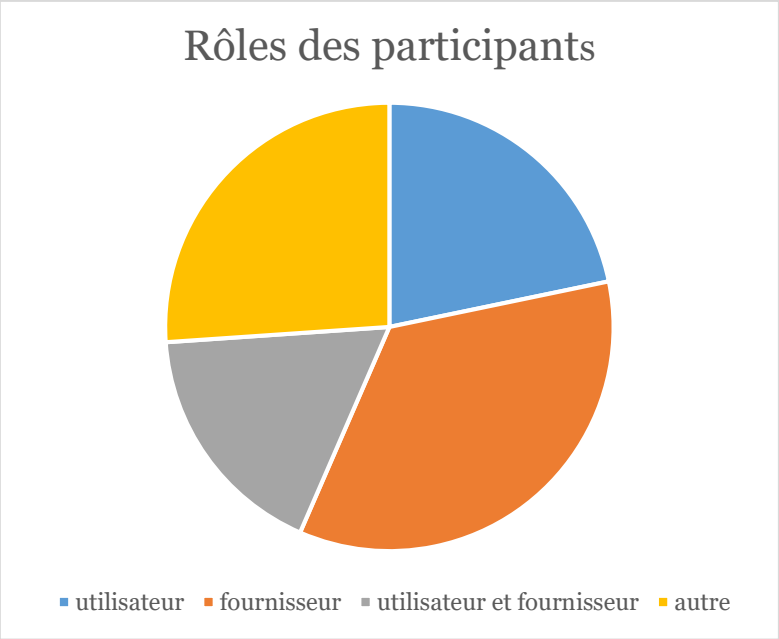


Les fonctions des personnes ayant répondu au questionnaire sont principalement juridiques à 59% (16 réponses), puis 22% ont un rôle exécutif (6 réponses), enfin, 19% ont un rôle technique lié au domaine de

l'intelligence artificiel (5 réponses). Aucune réponse n'a été rédigée par une personne ayant un rôle technique directement lié à la sécurité des systèmes d'informations.

Les rôles des participants au regard du règlement européen sur l'IA, étaient :

- à 30% celui de fournisseur de système (8 réponses) ;
- à 19% celui d'utilisateur (5 réponses) ;
- à 15% à la fois ceux de fournisseur et d'utilisateur (4 réponses) ;
- à 22% autre que fournisseur ou utilisateur (6 réponses).



2. Sur les risques de réidentification

Question 2.1

Quelles menaces peuvent conduire à une réidentification des personnes à partir du modèle entraîné ? Pour chacune de ces menaces, pourriez-vous décrire :

- **la source du risque (nature de l'attaquant s'il s'agit d'une source malicieuse, motivations, ressources financières et techniques, niveau d'accès aux données, au modèle, etc.) ;**
- **les objectifs visés par la source de risque.**

À cette première question, **toutes les réponses ont souligné l'existence de risques concernant la protection de la vie privée**.

Les risques mentionnés concernent certaines grandes catégories d'atteinte à la vie privée, telles que :

- L'usurpation d'identité ;
- le chantage et l'extorsion ;
- l'hameçonnage ;
- l'atteinte à la réputation.

Il a également été pointé que la possibilité d'extraire des données depuis un modèle entraîné pouvait être exploitée par un attaquant afin de nuire à une personne, par exemple en diffusant un modèle ayant mémorisé des informations compromettantes concernant un individu.
Les moyens techniques mentionnés permettant de réaliser ces risques sont :

- **les attaques par inférence d'appartenance** ;
- **les attaques par inférence d'attribut** ;
- **la régurgitation**.

Certaines indications supplémentaires ont été apportées concernant ces risques :

- Sur ces risques en général, il a été indiqué qu'ils peuvent être difficiles à mettre en œuvre en pratique car ils peuvent nécessiter des **compétences particulièrement rares** et avoir un **coût important**. Ainsi, seuls certains chercheurs spécialisés seraient à même de les mettre en œuvre. Elles nécessitent parfois d'avoir **accès à une portion du jeu de données** d'entraînement. Elles sont également **plus difficiles à mettre en œuvre lorsque le modèle n'est accessible qu'en boîte noire** (*black box*), alors qu'un contexte d'ouverture des poids les rend plus vulnérables.
- Il a également été soulevé qu'il était **difficile pour un attaquant d'identifier lorsqu'une attaque était un succès** puisque des données tierces sont généralement nécessaires pour vérifier la véracité de l'information obtenue. Ces attaques ne pourraient ainsi pas systématiquement être qualifiées de réidentification. Toutefois, l'accessibilité de données tierces, via des moteurs de recherche par exemple, pourrait faciliter ce recoupement.
- Concernant les attaques mises en œuvre par des entreprises concurrentes, il a été souligné que leur objectif visera davantage à **porter atteinte au développeur d'un modèle** qu'aux personnes concernées, par une **violation de la propriété intellectuelle** du jeu de données d'entraînement par exemple.

Une distinction a parfois été faite selon les catégories de modèles utilisés. Parmi les modèles peu risqués ont été cités les réseaux convolutifs (*convolutional neural network* ou CNN), les perceptrons multicouches, ou encore les transformeurs utilisé pour la vision par ordinateur (ViT). Les modèles tels que les grands modèles de langage et les modèles vision-langage ont été considérés comme à risque principalement lorsqu'ils sont génératifs. Des risques peuvent être plus importants que d'autres pour certains modèles, à l'instar des modèles de reconnaissance d'entité nommées, particulièrement sensibles aux attaques par inférence d'appartenance.

Les répondants ont identifié certaines sources de risques, tels que :

- les pirates informatiques, ou cybercriminels ;
- les utilisateurs finaux des systèmes d'IA génératifs ;
- les entreprises concurrentes ;
- les acteurs étatiques ;
- les employés mécontents ou mal intentionnés.

Lorsqu'il s'agit d'attaquants et non d'utilisateurs, les motivations citées sont l'espionnage industriel, la manipulation de décisions pour le bénéfice personnel, le gain financier par la demande d'une rançon ou le chantage.

Enfin, quelques scénarios spécifiques ont été développés dans les réponses :

- **Scénario 1.** Les modèles d'IA peuvent être exploités afin de retrouver des informations sur des personnes spécifiquement visées. La connaissance de ces informations par l'attaquant dans un certain contexte peut alors leur nuire. Le cas cité concerne l'utilisation d'un modèle afin de connaître les antécédents judiciaires d'un candidat dans un processus de recrutement. L'employeur peu scrupuleux utiliserait ainsi l'outil afin de discriminer certains candidats.
- **Scénario 2.** Un attaquant disposant d'ores et déjà d'information sur des personnes, comme leur numéro de téléphone ou leurs adresses mail pour exploiter ces informations afin d'extraire des données d'un modèle entraîné sur des historiques de service après-vente. Les informations obtenues sur les personnes, telles que leur historique d'achat, pourraient être utilisées afin de mieux cibler une campagne d'hameçonnage en se faisant passer pour le commerçant.
- **Scénario 3.** L'attaque par inférence d'appartenance pourrait être mise en œuvre par un attaquant afin d'obtenir des informations (directes ou non) sur l'état de santé d'un individu. Ainsi, un modèle visant à déterminer le meilleur contrat selon le profil d'un client pourrait permettre de connaître le contrat de l'une des personnes dont les données ont été utilisées pour l'entraînement, et ainsi révéler des clauses spécifiques à certaines pathologies ou risques sur la santé. Ce même type d'attaque pourrait être

appliqué à un modèle de langage tel qu'un *masked language model* entraîné sur des courriers médicaux, pourrait être exploité afin de connaître la pathologie d'une personne,

- **Scénario 4.** Un modèle permettant de prédire le montant du salaire d'une personne pourrait être attaqué pour déterminer le salaire d'une personne dont les données ont été utilisées pour l'entraînement.

Question 2.2

Quels facteurs peuvent avoir un impact (positif ou négatif) sur la mémorisation lors de l'entraînement du modèle (notamment les facteurs pouvant augmenter le risque de mémorisation, de régurgitation ou la facilité d'une attaque) ?

Les réponses apportées peuvent être réparties en trois catégories selon si elles portent **sur les données d'entraînement, le protocole d'apprentissage, ou encore sur le modèle lui-même**. Il est à noter que les réponses apportées ciblaient parfois les facteurs influençant les risques de régurgitation ou d'extraction de données, davantage que les facteurs influençant la mémorisation.

Les facteurs cités portant **sur les données d'entraînement** sont :

- **Le volume** de données d'entraînement au regard de la complexité du modèle (c'est-à-dire du nombre de paramètres).
- **Le nombre de doublons**, c'est-à-dire d'occurrence de données identiques ou similaires.
- **La présence de données rares**, c'est-à-dire positionnées en bordure de la distribution statistique du jeu d'entraînement.
- **Le caractère identifiant des données**, et en particulier les mesures prises afin de les pseudonymiser, de les anonymiser, ou encore l'utilisation de données synthétiques. Ces mesures incluent par exemple l'augmentation des données (modification mineure des données, comme un effet miroir ou un recentrage pour une image), la confidentialité différentielle ou la randomisation des données. Un répondant a également indiqué que l'existence d'un lien entre différentes données concernant une même personne (direct comme un identifiant, ou indirect comme un horodatage), ainsi que la proportion de données concernant une même personne pouvait augmenter le risque de mémorisation. Un répondant a toutefois précisé que la pseudonymisation était coûteuse, difficile à mettre en œuvre, et n'apportait pas de bénéfice additionnel dans le cas où les doublons avaient été supprimés par déduplication et où le volume de données était faible en comparaison du nombre de paramètres du modèle.
- **La qualité des données** entendue dans un sens large, et en particulier leur **diversité**.

Les facteurs cités portant **sur le protocole d'apprentissage** sont :

- **Le surapprentissage** ayant eu lieu lors de l'entraînement. Certaines techniques permettent toutefois de réduire son impact sur la mémorisation, comme le recours aux méthodes de régularisation, en particulier le *drop-out* (bien qu'une réponse ait indiqué que les preuves des effets du *drop-out* manquent encore), l'ajout d'une perturbation aléatoire à la fonction de coût, le choix de certaines fonctions de coût visant à éviter la mémorisation. Il a également été souligné que les preuves manquaient actuellement pour affirmer qu'une donnée dont l'influence sur les paramètres serait grande aurait plus de chances d'être extraite.
- **Les mesures d'anonymisation** prises lors du développement du modèle, comme la confidentialité différentielle lors de l'apprentissage, bien que cette technique ait un coût important (sur la performance du modèle et en temps de calcul), ou la perturbation des paramètres du modèle.

Les facteurs cités portant **sur le modèle** sont :

- **La fonctionnalité du modèle**, les modèles génératifs étant plus susceptible de mémoriser que les autres catégories de modèle.

- **La taille du modèle**, puisqu'un modèle de petite taille pourra être manipulé plus facilement par un attaquant, alors que les ressources nécessaires pour attaquer un grand modèle de fondation sont plus importantes. Ce facteur, ici considéré comme ayant un impact positif sur les risques d'attaque, est à mettre en balance avec l'impact négatif sur la mémorisation du choix d'un modèle plus petit (à quantité de données égales).
- **Les modifications apportées au modèle entraîné** visant à réduire la quantité d'information qu'il contient, comme la quantification de ses paramètres, la compression du modèle, ou l'élagage (ou *pruning*, bien qu'une réponse ait souligné que les preuves de l'effet de l'élagage manquaient encore). L'ajustement du modèle (*fine-tuning*) a également été cité parmi les techniques impactant négativement la mémorisation pour les données de l'entraînement principal, bien que cela entraîne un risque de mémorisation des données d'ajustement.

Question 2.3

Quels facteurs peuvent avoir un impact (positif ou négatif) sur la vraisemblance d'une régurgitation ou de l'extraction de données (notamment ceux portant sur les motivations de l'attaquant, sur la facilité à conduire l'attaque, l'existence d'attaques alternatives, etc.) ?

Les réponses apportées fournissent de manière complémentaire les facteurs pouvant impacter la vraisemblance d'une extraction des données d'entraînement ou d'une régurgitation, ainsi que les mesures à prendre afin d'influencer ces risques. Ces facteurs et mesures peuvent être classés en deux catégories : ceux portant sur **la motivation de l'attaquant**, et ceux impactant **les chances de succès d'une attaque ou la vraisemblance d'une régurgitation**. A noter que la régurgitation pouvant avoir lieu dans le cadre de l'utilisation normale d'un modèle génératif, les facteurs portant sur la motivation de l'attaquant manqueront de pertinence dans ce cas.

Les facteurs et mesures cités impactant **la motivation de l'attaquant** sont :

- **La nature des données d'entraînement et leur intérêt pour un potentiel attaquant**, tel que le fait qu'il s'agisse de données publiquement accessibles. Ce facteur étant souvent lié à la sensibilité des données, et ainsi aux potentielles conséquences pour les personnes. Les mesures concernant la sélection et la pseudonymisation des données d'entraînement peuvent contribuer à réduire la motivation de l'attaquant si ce dernier en a connaissance. Ces mesures sont listées plus bas parmi les mesures impactant les chances de succès d'une attaque.
- En lien avec la nature des données d'entraînement, **la nature de l'organisme susceptible d'être attaqué** peut également influencer la motivation de l'attaquant.
- **La proportion de données et la nature de l'information pouvant être obtenue** via une attaque. Le résultat d'une attaque dépend en effet des contextes étudiés, puisqu'une attaque par inférence d'appartenance peut n'apporter aucune information utile à un attaquant dans certains contextes, ou au contraire révéler des informations sensibles.
- **L'existence de surfaces d'attaque alternatives**, telles qu'un accès à la base de données d'entraînement. Cet accès peut être protégé, mais outrepasser les sécurités d'accès peut parfois être moins complexe que d'extraire les données du modèle.
- **Certaines réponses ont souligné le coût (financier et temporel) et la complexité technique** des attaques considérées et pertinentes selon le contexte. Ainsi, le temps nécessaire pour la génération de sorties via une API, ou encore le coût de l'infrastructure et du matériel nécessaires pour utiliser et attaquer un modèle ouvert peuvent accroître le coût d'une attaque. De même, l'existence d'une documentation permettant de reproduire une attaque peut la rendre plus vraisemblable.

Les facteurs et mesures cités impactant **les chances de succès d'une attaque ou la vraisemblance d'une régurgitation** sont :

- Concernant les données :
    - **L'accès, au moins partiel, à des connaissances sur les données d'entraînement.** Ces informations peuvent inclure les données elles-mêmes, une description du jeu utilisé, ou encore

le fait que les données de la personne ciblée par l'attaque sortent de la distribution statistique de l'ensemble du jeu (données rares ou *outliers*).

- **La proportion de données identifiantes dans le jeu d'entraînement**, et, en parallèle, les mesures de pseudonymisation ou d'anonymisation mises en œuvre. Ces mesures incluent par exemple l'utilisation de données synthétiques en remplacement ou en complément de données personnelles, l'attention portée à la collecte, la sélection des données, le masquage ou le remplacement automatique ou manuel des informations, et en particulier de certaines catégories (*feature engineering*), ou encore l'utilisation de techniques apportant des garanties de confidentialité différentielle. La pertinence de ces mesures a toutefois été questionnée en raison de leur coût, de leur complexité et de leur impact sur la performance des modèles entraînés. Le choix des données spécifiquement utilisées pour l'entraînement permet également de limiter la quantité de données personnelles dans les données d'entraînement. Cette sélection peut être facilitée en labellisant les données en amont (*tagging*) et par une stratégie de regroupement des données rigoureuse (*pooling*).
- **Le volume de données au regard de la complexité du modèle**, quantifiée par son nombre de paramètres.
- **Le nettoyage des données**, et notamment la suppression des outliers et des doublons (déduplication).

- Concernant le modèle :
  - **Les conditions d'accès au modèle** peuvent faciliter une attaque, notamment lorsqu'il s'agit d'un modèle dont les poids sont ouverts, ou en source ouverte. Au contraire, attaquer les modèles à accès restreint, ou accessibles via une API sécurisée est plus complexe, notamment lorsque les « logits » liés à l'inférence ne sont pas fournis.
  - **L'accès à des informations tierces**, comme d'autres versions du modèle (notamment pour l'apprentissage en continu, ou les versions antérieures du modèle peuvent renseigner sur les dernières données d'entraînement), ou encore à d'autres modèles au moins partiellement entraînés sur les mêmes données d'entraînement. La connaissance de la stratégie d'apprentissage ou d'alignement pour les grands modèles de langage peut également faciliter les attaques.
  - **L'accès à la documentation** sur le modèle peut également faciliter les attaques ou au contraire les limiter, au détriment de la transparence.
  - **Le type de modèle considéré**, notamment selon s'il est génératif ou prédictif.
  - **Les mesures visant à éviter la mémorisation** lors de l'apprentissage, comme la régularisation, le *drop-out*, l'ajout d'aléatoire dans les paramètres du modèle, ou les techniques de de confidentialité différentielle comme le protocole PATE (*Privacy Aggregation of Teachers Ensembles*).
  - **Les modifications apportées au modèle**, comme l'élagage (*pruning*), la quantification de ses paramètres ou sa compression, l'ajustement du modèle, par exemple par l'apprentissage par renforcement à partir de la rétroaction humaine (*reinforcement learning from human feedback*, ou RLHF). Une réponse a indiqué que les techniques de désapprentissage machine manquaient encore de maturité.
- Concernant le système d'IA :
  - **La modalité de la mise à disposition ou du déploiement du système**, y compris le nombre d'utilisateurs prévu, la nature des utilisateurs (grand public, entreprise privée, etc.), la sécurité du réseau utilisé, le déploiement en nuage, ou local.
  - **Le cadrage des requêtes**, la limitation du taux de requêtage et l'utilisation de filtres (couramment appelés *guardrails*) sur les requêtes envoyées au système visant à identifier et bloquer les entrées malveillantes, et notamment les tentatives de débridage (*jailbreaking*). Une fois les attaques identifiées, des mesures peuvent être prises à l'encontre de l'attaquant, comme le blocage de compte ou le retrait du service. Des mesures sont également mentionnées afin de désamorcer les tentatives d'attaque, ou d'éviter les régurgitations, comme l'utilisation d'une pré-requête cadrant l'inférence, ou la contextualisation des requêtes (par le *Retrieval Augmented Generation*, ou RAG, par exemple).
  - **Les limitations et filtres sur les sorties** réduisant la quantité d'information pouvant être révélée (comme les « logits »), ou ciblant certaines informations protégées sont citées. L'analyse des réponses, par des techniques de traitement automatique du langage par exemple, permet également de cibler, voire de modifier les réponses contenant des données personnelles. Les

mesures visant à éviter les hallucinations, comme la limitation de la longueur des réponses, peuvent également permettre d'éviter les régurgitations.

- Concernant le résultat de l'attaque :
    - **La possibilité de vérifier la véracité des informations obtenues**, par le biais de connaissance annexes par exemple, ou la possibilité pour l'attaquant de vérifier cette véracité (de manière automatisée ou non, par exemple avec une requête pour une adresse IP, ou en tentant d'appeler directement un numéro de téléphone).

À noter que ces réponses recoupent en grande partie ceux ayant un impact sur la vraisemblance d'une mémorisation.

Enfin, certaines mesures plus générales ont été mentionnés, telles que :

- La réalisation d'une analyse de risques, via l'utilisation de la méthode EBIOS notamment, des audits de sécurité, ou une comparaison avec un modèle similaire entraîné sur des données différentes.
- le chiffrement du modèle, afin de limiter la possibilité de l'exploiter lors d'une attaque en cas de divulgation**,**
- La création d'une politique de respect de la vie privée en interne pour insuffler un comportement vertueux et améliorer les pratiques.

Question 2.4

Quelles pourraient être les conséquences d'une régurgitation ou d'une extraction de données pour les personnes dont les données ont été utilisées pour l'entraînement ?

Parmi les réponses obtenues, plusieurs ont indiqué que ces conséquences étaient les mêmes que celles habituellement rencontrées pour toute divulgation de données personnelles. L'ampleur des cas d'usages et des catégories de données pouvant être traitées pour l'entraînement de systèmes d'IA a conduit certains participants à considérer qu'**une grande diversité de risques pouvait s'appliquer,** comme les pressions politiques, les cambriolages, le harcèlement en ligne, ou encore le refus de l'attribution d'un contrat ou d'une assurance. Certaines des conséquences mentionnées ne portaient pas sur un individu en particulier, comme le sentiment généralisé de perte du droit à la vie privée, au contraire, un faux sentiment de sécurité (les données étant sécurisées, mais les modèles non), ou encore la perte de confiance dans les techniques et l'écosystème de l'intelligence artificielle.

Certaines réponses ont toutefois ciblé certaines conséquences en particulier, comme :

- l'atteinte à la réputation, avec des conséquences particulièrement importantes lorsque la régurgitation est liée à une hallucination (c'est-à-dire pour la génération de fausses informations sur un individu réel) ;
- l'usurpation d'identité (notamment pour la fraude financière) ;
- l'hameçonnage et les arnaques ;
- la discrimination ;
- le non-respect de la propriété intellectuelle **;**
- le démarchage à des fins marketing ;
- la manipulation ***;***
- la perte de contrôle sur ses données, et en particulier la difficulté à exercer ses droits.

Question 2.5

Quels facteurs peuvent avoir un impact (positif ou négatif) sur la gravité des conséquences pour les personnes ?

Les réponses fournies listent des facteurs portant **sur les données d'entraînement, les personnes concernées, ou le contexte de l'attaque**.

Les facteurs portant sur **les données d'entraînement** sont :

- **La nature des données** d'entraînement, et notamment s'il s'agit de données sensibles ou hautement personnelles, ou de données publiquement accessibles ou inexploitables par un attaquant. Sur ce sujet, une réponse a soulevé que certaines informations a priori anodines, comme un historique d'achat, peuvent être exploitées par un attaquant afin d'obtenir la confiance de la personne concernée et de la tromper. Les mesures de pseudonymisation peuvent également permettre de limiter les conséquences d'une divulgation,
- **La proportion de données personnelles dans la base d'entraînement**, et la quantité d'informations portant sur un même individu.
- **L'exactitude des données** d'entraînement, ainsi que leur actualité, pouvant autant diminuer la gravité des conséquences quand elle empêche la réidentification, ou l'augmenter quand les informations portent atteinte à la réputation de la personne,
- **La maîtrise sur les données** d'entraînement et sur leur contenu,
- **La réutilisation des données** des utilisateurs pour ajuster un modèle initialement entraîné sur des données publiquement accessibles.

Sur **les personnes concernées**, une réponse a mentionné que ces conséquences pouvaient être particulièrement graves pour certaines des personnes concernées, notamment en raison du caractère indiscriminé de la collecte de données d'entraînement pour l'IA, qui peut notamment toucher des enfants.

Les facteurs portant sur **le contexte de l'attaque** sont :

- **L'existence d'informations tierces**, éventuellement fournies par des systèmes d'IA différents, pouvant être mises en relation avec les données obtenues.
- **L'incertitude sur la véracité des informations obtenues** et les moyens permettant à l'attaquant de les vérifier.
- **La visibilité de l'incident** et des données confidentielles divulguées pouvant notamment être augmentée par le nombre d'utilisateurs du système.
- La facilité à conduire les attaques puisque celle-ci peut amener l'attaquant à extraire davantage de données (cela peut être le cas lorsque la surface d'attaque est importante, ou quand les chances de succès d'une attaque sont accrues).
- **La préparation et la sensibilisation des individus** aux risques liés à la divulgation des données et les moyens de recours prévus.
- **La formation des développeurs** ayant conçu le modèle à ces risques et la réactivité de l'organisme en cas de violation de données personnelles (par exemple à mettre en œuvre une information des personnes concernées, à retirer le modèle ou à le modifier).
- Le fait qu'il soit impossible d'exercer certains droits, comme le droit à l'oubli, sur le modèle.
- **Le type de système sujet à l'attaque**, et notamment la possibilité qu'il ait été utilisé pour des usages détourné dans le cadre desquels les utilisateurs ont dévoilé des informations plus sensibles que celles attendues dans le contexte d'utilisation normal du système.

À noter qu'une réponse a soulevé que certains des facteurs listés par la CNIL, tels que le nombre d'attaquant, ou la sensibilité des données, n'ont pas d'impact lorsque l'attaquant ne peut être certain de la véracité des informations obtenues.

Question 2.6

Dans l'hypothèse où l'analyse du risque de régurgitation ou d'extraction portant sur le modèle ne serait exigée que dans certains cas, une liste de critères aidant à identifier ces situations semble nécessaire. La liste suivante pourrait permettre de déterminer les cas où les risques de régurgitation et d'extraction sont suffisamment faibles (ces cas sont caractérisés par le fait qu'aucun des critères ci-dessous n'est rempli). Les critères de cette liste vous semblent-ils pertinents ? Cette liste vous semble-t-elle exhaustive ?

- **Le caractère identifiant des données, comme la présence de noms, prénoms, photographies du visage, adresses ou dates de naissance dans le jeu d'entraînement ;**
- **La présence de données rares ou aberrantes, aussi appelées *outliers* ;**
- **La duplication des données d'apprentissage ;**
- **Le nombre de paramètres du modèle au regard du volume de données ;**
- **La fonctionnalité du modèle (générative ou prédictive par exemple) ;**

Le surapprentissage (par exemple une métrique montrant une meilleure performance sur les données d'entraînement que sur les données de validation).

Plusieurs participants ont listé les critères jugés pertinents, permettant d'en réaliser un classement :

- le caractère identifiant des données ;
- la fonctionnalité du modèle (générative ou prédictive par exemple) ;
- le nombre de paramètres du modèles au regard du volume de données ;
- le surapprentissage ;
- la duplication des données d'apprentissage ;
- la présence de données rares ou aberrantes.

Les réponses ont ciblé les critères suivants, jugés pertinents et pouvant être ajoutés à cette liste :

- la finalité du modèle (un modèle utilisé pour authentifier des personnes par exemple a été jugé plus à risque) ;
- les modalités d'accès au modèle, et notamment le niveau d'ouverture du modèle et l'échelle de son déploiement ;
- l'interopérabilité du système, lui permettant de recouper avec des informations provenant de sources tierces ou d'autres systèmes d'IA ;
- la proportion de données sensibles dans le jeu d'entraînement ;
- la maîtrise sur les données d'apprentissage et notamment la connaissance de leur contenu et la fiabilité des sources ;
- pour l'IA générative, la permissivité des requêtes, et les mesures de sécurité qui leur sont apportées ;
- l'existence d'une littérature sur les risques d'attaque pour le modèle concerné, et ainsi la vulnérabilité connue du modèle aux attaques ;
- la proportion de données personnelles, et notamment de données sensibles dans le jeu ;
- la transparence et l'explicabilité du modèle, puisque la possibilité d'expliquer les prédictions d'un modèle peut aider à évaluer si les sorties reposent sur des caractéristiques sensibles ou identifiables des données ;
- la diversité des données utilisées qui tend à réduire les mémorisations excessives ;
- l'utilisation de méthodes d'anonymisation comme la confidentialité différentielle, ou de "modèles abstraits" dont le niveau de pseudonymisation est démontré par des méthodes empiriques (comme l'agrégation ou la distillation) ;
- le protocole d'apprentissage, et notamment l'ajustement et l'apprentissage par renforcement sur des données personnelles.

Au contraire, la pertinence de certains critères a été questionnée, comme :

- La présence d'identifiants dans les données, celle-ci ne présentant un risque que lorsque les identifiants permettent de relier des données d'une même personne entre elles. Une autre réponse indique que cette simple présence n'est pas un critère suffisant, et devrait être complété par une étude de la nature des données et de la quantité des données par personne.

- La duplication des données d'entraînement, qui concerne de trop nombreux cas et peut être complexe à mesurer alors que la mémorisation peut concerner des données sans doublon.
- Le surapprentissage, puisque celui-ci aura systématiquement lieu dans une certaine mesure, et puisque son lien a la mémorisation n'a pas été prouvé.
- La pseudonymisation, qui n'aurait pas d'impact en absence de duplication des données.
- La présence de données rares ou aberrantes, qui n'est pas un critère discriminant puisque cela sera souvent le cas, alors que ces données ne sont pas toujours des données personnelles et que leur présence peut être nécessaire à l'apprentissage. De plus la notion de donnée aberrante n'est pas applicable dans tous les cas, et notamment pour les données de hautes dimensions comme les images.
- Le nombre de paramètres du modèle au regard du volume de données, car le volume de données n'est pas toujours représentatif de la quantité d'informations contenues dans le jeu. De plus, ce critère n'entraîne pas toujours un risque accru de régurgitation ou d'extraction.

Plusieurs réponses ont ainsi soulevé qu'en l'absence de seuil fixé sur ces critères, la liste pourrait être trop exhaustive et impliquer qu'une analyse approfondie serait nécessaire pour tout modèle entraîné sur des données personnelles. Une réponse a également souligné que les modèles d'apprentissage profond ne sont pas nécessairement les plus à risque puisque leur complexité apporte de la difficulté pour l'attaquant.

3. Sur les techniques permettant d'analyser les risques de régurgitation et d'extraction de données personnelles

Question 3.1

Quelles sont les techniques pertinentes pour l'analyse des risques de régurgitation et d'extraction de données ? Ces mesures peuvent chercher à évaluer le risque de mémorisation par le modèle, de régurgitation ou d'une attaque spécifique (par inférence d'appartenance, reconstruction, etc.).

Parmi les réponses obtenues, de nombreuses ont pointé vers la méthodologie de l'équipe adverse (*red teaming*) dans le cadre de laquelle une équipe tente d'extraire des données à partir du modèle entraîné au moyen de diverses techniques. Toujours d'un point de vue méthodologique, une réponse a indiqué que cette analyse devait s'inscrire dans l'analyse de risques prévue par le RGPD, et une autre qu'elle devait mesurer l'impact social des régurgitations et extractions. Une troisième réponse a pointé que l'analyse devait reposer sur des *benchmarks* reconnus comme PrivQA. Il a été suggéré que les tests réalisés portent sur les données les plus à risque comme les données rares ou aberrantes (*outliers*), et les données les plus susceptibles de motiver un attaquant. Une réponse a indiqué que des tests devaient également porter sur les systèmes qui reposent sur le modèle, dont les fonctionnalités peuvent augmenter la probabilité d'une attaque ou la réduire.

D'un point de vue technique, les méthodes citées pour conduire l'analyse sont :

- Les tests portant sur les données : une proportion importante de données personnelles indique un risque plus important, en particulier si elles sont susceptibles de motiver un attaquant.
- L'utilisation de techniques de confidentialité différentielle, qui permettent de vérifier la confidentialité des données d'entraînement et parfois du modèle entraîné.
- L'utilisation d'attaques fictives par inférence d'appartenance, dont la première étape peut constituer à mesurer le surapprentissage, en notant qu'en cas de réussite, la procédure mise en œuvre permet de mesurer le coût et la complexité de l'attaque.
- Lorsqu'une interface avec le système existe, l'utilisation de requêtes spécifiques contenant par exemple des mots-clés se rapportant aux données d'entraînement.
- Les tests visant à évaluer le risque de débridage du système (*jailbreak*), par exemple en tenant de reconstituer la pré-requête (*preprompt*) visant à éviter ces attaques.
- Les tests de pénétration dans le système dans son ensemble, afin d'évaluer la sécurité des points d'accès au modèle.

- Pour un modèle de langage masqué (*masked language model*) spécifiquement, les tests reposant sur la génération de mots spécifiques dans les textes du jeu d'entraînement.
- L'approche contrefactuelle, qui repose sur l'entraînement d'un second modèle sur le même jeu d'entraînement privé des données à protéger : lorsque les performances du modèle diffèrent, cela révèle que les données protégées ont joué un rôle.
- Les tests visant à évaluer le caractère reproductible ou généralisable des attaques, puisque leur caractère probabiliste d'une part, et la possibilité de les appliquer à une grande proportion de données d'autre part, influencent la vraisemblance et la gravité de ces failles.

Malgré l'existence de ces techniques, plusieurs réponses ont indiqué qu'elles semblaient insuffisantes actuellement pour conduire une analyse robuste et qu'aucun consensus n'existait sur la méthodologie à adopter. Une autre réponse a indiqué que la plupart des techniques de test documentées actuellement portaient sur les grands modèles de langage, au détriment des autres catégories de modèles. Une réponse a ainsi suggéré qu'un répertoire tel que le *Common Vulnerabilities and Exposures*[1] appliqué aux risques sur les modèles permettrait d'avancer vers l'élaboration d'un standard pour cette analyse.

Enfin, certains participants ont soulevé qu'avec la mise en œuvre de mesures en amont, comme la déduplication des données, l'utilisation de données publiquement accessibles et un nombre de paramètres suffisamment petit en comparaison du volume de données, l'analyse ne semble pas nécessaire.

Question 3.2

Un seuil sur le risque de régurgitation ou d'extraction de données tel que mesuré grâce aux techniques listées à la question précédente peut-il être fixé afin de considérer le traitement du modèle doit être regardé, en lui-même, comme un traitement de données à caractère personnel ? Comment fixer ce seuil ? D'autres critères vous semblent-ils pertinents ?

Les réponses à cette question se placent dans deux catégories :

- celles soutenant qu'un seuil ne peut pas être fixées : elles sont au nombre de 5 ;
- celles soutenant que ce seuil peut exister : elles sont au nombre de 4.

Les participants soutenant qu'un tel seuil ne peut être fixé ont appuyé leur réponse sur le fait que les techniques d'évaluation manquent encore de robustesse, et que ces techniques – et ainsi le seuil correspondant – dépendaient d'une part du modèle évalué et d'autre part de la quantité de données personnelles contenues dans les données d'entraînement. Il a également été soulevé que les praticiens manquent encore de connaissances sur les techniques d'attaque, et que le seuil fixé risquerait ainsi de ne pas être correctement appliqué.

Les réponses indiquant qu'un seuil pouvait être fixé ont suggéré la piste de la confidentialité différentielle, ou encore celle d'un taux maximum de succès pour les attaques. Ce seuil pourrait comporter plusieurs critères comme la déduplication et une comparaison du nombre de paramètres du modèle au regard du volume de données d'entraînement. Deux réponses ont en effet suggéré que le seuil porte sur le résultat de l'analyse des données d'entraînement plutôt que sur celle du modèle. Il a été pointé que pour certaines catégories de modèles, les techniques d'attaque sont aujourd'hui suffisamment stabilisées pour fixer une méthodologie et un cadre d'analyse.

Dans l'éventualité où un seuil pourrait être fixé, certaines réponses ont indiqué que ce seuil devrait être adapté en fonction des cas d'usage, des modalités d'accès au modèle (du moyen d'accès, comme les API, et de la fréquence de requêtage du modèle), du type de modèle, de la quantité et de la nature des données personnelles (et ainsi des conséquences pour les personnes). En effet, il a été pointé que les seuls risques de régurgitation et d'extraction ne pouvaient suffire à qualifier le modèle, car les risques proviennent de sources externes dont les moyens semblent encore insuffisants. La pertinence d'un tel seuil a été questionnée, dans la mesure ou il ne tiendrait pas compte des mesures complémentaires mises en œuvre et de la vraisemblance d'une attaque.

[1] https://cve.mitre.org/

Enfin, il a été pointé que ce seuil ne semblait pas nécessaire, pour deux raisons bien distinctes. D'une part certaines réponses ont soutenu que l'utilisation de critères comme le fait que le modèle soit génératif était préférable à l'utilisation d'un seul numéraire. D'autre part, certaines réponses ont avancé que le modèle ne pouvait être soumis au RGPD, et ainsi que cette question ne pouvait porter que sur les traitements de développement ou de déploiement.

Question 3.3

Parmi ces techniques, identifiez-vous des difficultés quant à leur mise en œuvre ?

Parmi les difficultés mentionnées, **certaines sont d'ordre organisationnel**. Les réponses indiquaient qu'elles nécessitaient notamment une formation importante des délégués à la protection des données d'une part, et des experts techniques d'autre part. Il a également été largement soulevé que ces techniques nécessitaient un investissement en temps et en ressources important, en particulier pour les acteurs qui doivent actuellement se mettre en conformité avec le règlement IA. Bien que des réponses aient indiqué que l'analyse portant sur les données était possible, car les techniques sont plus simples, connues et documentées, et car elles peuvent parfois être automatisées. Ces mesures représenteraient toutefois un coût et une complexité importante pour les grands jeux de données non structurées. Il a aussi été soulevé que si l'obligation de l'analyse reposait sur le concepteur, les déployeurs pourraient être désincités à évaluer les risques pour leur cas spécifique. Lorsque l'analyse est effectuée par des tiers, comme des chercheurs ou la communauté *open source*, les restrictions d'accès au modèle et aux données complexifient l'analyse.

D'un point de vue technique et méthodologique, l'absence de consensus, et en particulier de standards sur la méthodologie et d'outils a également été pointée. Le manque de robustesse et de fiabilité des techniques, qui pourrait être la cause de cette absence, est également un frein à leur mise en œuvre. L'efficience de l'analyse a également été questionnée puisque certains concepteurs pourraient optimiser leurs modèles pour passer les tests (notamment si les tests sont publics et manquent de variété). La nécessité de mettre à jour l'analyse pour les nouvelles versions des modèles, et pour tenir compte de l'évolution des techniques d'apprentissage et d'attaque représente une autre difficulté. La nécessité d'utiliser des tests spécifiques à chaque type de modèle rend la généralisation de l'analyse complexe, en particulier car les tests les plus courants (comme l'inférence d'appartenance) ne sont pas applicables dans tous les cas. L'analyse devrait également être effectuée pour différentes implémentation du modèle pour tenir compte de l'impact de l'infrastructure et du cas applicatif sur les risques. Le seuil permettant d'acter que le niveau de risque serait suffisamment faible a aussi interrogé, notamment pour la probabilité de régurgiter des données concernant un individu, puisque ce résultat est parfois souhaité pour les personnalités publiques.

Question 3.4

Les techniques que vous avez recensées ou évoquées plus haut en exemple vous semblent-elles suffisantes pour évaluer les risques de régurgitation et d'extraction de données ? Permettent-elles de déterminer quand ces risques sont nuls ou suffisamment faibles, c'est-à-dire quand un modèle est anonyme ? Merci de justifier.

À cette question, une partie importante des réponses étaient négatives. Il a été soulevé en particulier que les techniques d'analyse n'étaient pas suffisamment fiables actuellement et encore en cours d'évolution. De plus, puisqu'elles reposent sur des attaques fictives, elles peuvent aider à démontrer l'existence d'un risque mais ne permettent pas de garantir son absence. De plus, il a été pointé que ces mesures étaient exigeantes pour les entreprises de petite taille. Les mesures d'anonymisation sur les jeux de données ont semblé plus pertinentes à certains participants. De même, une réponse a indiqué que l'analyse de risque devrait porter principalement sur les données d'entraînement. Enfin, une réponse a objecté que l'analyse n'était pas pertinente car les modèles d'IA n'étaient pas des objets auxquels le RGPD pouvait s'appliquer.

D'autres réponses ont indiqué que bien que ces techniques ne soient pas infaillibles, il semblait possible de s'y fier pour qualifier le modèle, au risque de devoir changer cette qualification avec l'apparition d'une faille. Ainsi, la méthodologie d'analyse devrait être mise à jour régulièrement avec l'évolution des techniques. Des

participants ont indiqué que ces techniques devaient être complétées par les mesures de sécurité techniques et organisationnelles habituelles.

Question 3.5

Dans quelles situations vous semble-t-il nécessaire de reconduire l'analyse d'un modèle ?

Les participants ont considéré que l'analyse devrait en premier lieu être reconduite lors d'une modification substantielle du modèle. Ces modifications incluent le réentraînement, l'ajustement (*fine-tuning*), l'apprentissage par transfert, ou un changement dans l'architecture du modèle. Une nouvelle analyse a semblé particulièrement pertinente lors qu'une modification du modèle est réalisée après un changement dans les données personnelles contenues dans le jeu d'entraînement, ou si la distribution statistique des données a évolué (faisant apparaître de nouveaux *outliers* par exemple).

L'observation d'un surapprentissage lors de ces étapes devrait notamment inciter à répéter l'analyse. Dans le cas de l'apprentissage continu, cette analyse devrait être répétée fréquemment, en particulier en raison du manque de visibilité sur les données d'entraînement qui peuvent contenir des données personnelles sans que ce soit attendu. Les changements dans le contexte de collecte des données d'entraînement (comme un changement dans les conditions d'utilisation, ou une évolution des pratiques, avec pour exemple le moissonnage de données sur Reddit de plus en plus utilisé comme messagerie privée). Lorsque des techniques de *retrieval augmented generation* (RAG) sont utilisées, l'analyse devrait être effectuée lors de modifications apportées aux données utilisées pour le RAG. Une réponse a pointé que l'analyse devrait être mise à jour tous les trois ans pour les traitements dont le niveau de risque est important.

Le contexte général devrait également être pris en compte d'après les participants. L'apparition de nouvelles méthodes d'attaque a été mentionnée à plusieurs reprise dans les réponses, ainsi que l'évolution des techniques améliorant la confidentialité. Le signalement d'une régurgitation par les utilisateurs, ou l'apparition d'une faille de sécurité devrait également motiver une nouvelle analyse. Enfin, plusieurs réponses ont indiqué que les changements dans les modalités d'accès au modèle (comme une modification dans les API), l'utilisation du modèle pour une autre tâche, ou une modification des mesures protectrices apportées au niveau applicatif pouvaient justifier une mise à jour de l'analyse.

Au contraire, une réponse a indiqué que seul l'ajustement sur des données personnelles pouvait justifier une nouvelle analyse. Enfin, une réponse a souligné que l'analyse n'était jamais pertinente, les modèles d'IA n'étant par nature pas des objets auxquels le RGPD pouvait s'appliquer.

4. Sur l'application du RGPD à un modèle d'IA ayant mémorisé des données personnelles

Question 4.1

S'il est établi qu'un modèle a mémorisé des données personnelles, à quelles conditions la simple détention du modèle doit-elle être assimilée à une conservation de données à caractère personnel soumise au RGPD :

- **uniquement lorsqu'il s'agit d'un modèle d'IA générative susceptible de régurgiter les données lorsqu'il est interrogé (*prompt*) ?**
- **y compris lorsque les données ne sont pas régurgitées dans le fonctionnement normal du modèle mais qu'elles peuvent être extraites par une attaque (inversion du modèle, attaque par inférence…) ?**

Selon d'autres critères ? Merci de justifier.

Plusieurs participants ont **supporté le principe d'une application du RGPD aux modèles** dès lors qu'une régurgitation ou une extraction de données personnelles contenues dans son jeu d'entraînement est possible. Dans le détail des réponses, les positions s'affinent et plusieurs pistes sont proposées :

- Une réponse suggère que le RGPD s'appliquerait par défaut s'il n'est pas possible de démontrer l'absence de risque de régurgitation ou d'attaque, en corollaire, le RGPD ne s'appliquerait plus lorsqu'il est démontré qu'il n'est pas possible d'inférer des données personnelles à partir du modèle. Une réponse précise que l'analyse permettant d'exclure l'application du RGPD devrait également tenir compte des protections comme les filtres évitant la régurgitation et de la sécurité du système.
- Une distinction est proposée entre les modèles génératifs et les autres catégories de modèles. Pour les modèles génératifs, le RGPD s'appliquerait lorsqu'il peut y avoir une régurgitation, alors que la nature des données traitées, la sécurité du déploiement et les modalités de mise à disposition du modèle devraient être prises en compte pour les autres catégories de modèles dont des données peuvent être extraites.
- Un participant indique que le RGPD ne devrait pas s'appliquer à un modèle ayant mémorisé lorsque son architecture ne permet pas la régurgitation.
- Une autre réponse suggère que la possibilité d'extraire des données personnelles des modèles entraînés ne peut justifier l'application du RGPD, car celles-ci sont trop exigeantes actuellement pour un attaquant, et leur résultat est incertain.
- Une réponse propose que le détenteur en aval du modèle soit exempté de l'application du RGPD lorsqu'il ne pouvait avoir connaissance du risque d'extraction de données personnelles depuis le modèle qu'il utilise.
- Un participant a indiqué que d'autres critères que le type de modèle devaient être pris en compte, comme la nature des données, les mesures de pseudonymisation des données, ou visant à éviter la mémorisation, ainsi que la vraisemblance et la gravité d'une attaque potentielle.

À l'opposé, certaines réponses ont considéré que **l'application du RGPD aux modèles ne peut reposer sur cette analyse** :

- Une réponse a considéré qu'en l'absence d'une méthodologie d'évaluation et de seuils bien définis, le RGPD s'appliquerait dans de nombreux cas où un faible risque de réidentification existe, sans conséquences réelles pour les personnes, impactant notamment le domaine de l'*open source*. La charge pourrait ainsi porter davantage le développeur du modèle qui aurait à fournir une documentation sur les risques pour la vie privée.
- Certaines réponses ont suggéré que la charge qu'implique cette analyse devrait être prise en compte, en particulier pour les TPE et PME déployeuses de systèmes.
- Une réponse a considéré que les mesures prises en amont, comme la déduplication et le faible nombre de paramètres au regard du volume de données d'entraînement, pouvaient exempter le développeur d'une analyse.
- Certaines réponses ont considéré que la position de l'autorité de protection des données de Hambourg dans sa publication « *Discussion Paper: Large Language Models and Personal Data* »[2] était adaptée et ainsi que le RGPD ne s'appliquait pas aux modèles,
- Enfin, une réponse a détaillé qu'en raison de l'absence de lien entre les paramètres d'un modèle et les personnes dont les données sont contenues dans le jeu d'entraînement, de l'incertitude sur les données obtenues par régurgitation ou par extraction et du travail nécessaire pour reconstruire des données personnelles à partir du modèle, la détention d'un modèle ne pouvait s'apparenter à une conservation de données personnelles. Le RGPD ne pourrait ainsi pas leur être appliqué.

5. Sur la responsabilité des acteurs

Question 5.1

Dans l'hypothèse où l'utilisateur aurait à conduire l'analyse du caractère anonyme du modèle, de quelles informations et ressources aurait-il besoin ? En dispose-t-il généralement selon

[2] https://datenschutzhamburg.de/fileadmin/user_upload/HmbBfDI/Datenschutz/Informationen/240715_Discussion_Paper_Hamburg_DPA_KI_Models.pdf

vous et, à défaut, qui les possède ? Vous semble-t-il possible pour l'utilisateur de réaliser l'analyse du caractère anonyme du modèle ?

Une part importante des réponses a considéré que l'utilisateur ne serait pas en mesure de réaliser l'analyse, sauf dans certaines situations. La réalisation de l'analyse par l'utilisateur demanderait une coordination importante avec le développeur du modèle, et qu'il lui fournisse par exemple :

- un accès aux données d'entraînement (notamment pour vérifier la véracité de l'information obtenue lors d'une attaque), à un échantillon, ou à une description (détaillant par exemple les catégories de données personnelles traitées, leur quantité au total et par individu, et les mesures de pseudonymisation prises) ;
- le code utilisé pour entraîner et exécuter le modèle ;
- une documentation technique du modèle, et du protocole d'apprentissage (indiquant notamment l'ordre dans lequel les données ont été traitées lors de l'entraînement et les mesures de régularisation utilisées) ;
- une description de la sécurité du système intégrant le modèle (comme ses vulnérabilités et les mesures de protection prévues).

Des protocoles semblent nécessaires pour faciliter ces échanges. Cette coordination pose toutefois certaines difficultés, notamment pour le partage de données protégées par la propriété intellectuelle ou par le RGPD. Un participant a suggéré que le fournisseur ne transmette que les informations nécessaires à la vérification des résultats de l'analyse qu'il aura conduite. Dans le cas où l'utilisateur aurait réalisé un ajustement sur des données personnelles, il pourrait alors conduire l'analyse. Une distinction pourrait également être faite pour les modèles prédictifs, pour lesquels davantage de transparence existe entre fournisseur et déployeur. Au contraire, les tests sur les modèles génératifs pouvant être mis en œuvre plus facilement, au moyen de requêtes par exemple, ces modèles pourraient être analysés plus facilement par les utilisateurs.

D'autres participants ont considéré que seul le fournisseur serait en mesure de réaliser l'analyse. L'accès à une puissance de calcul importante, aux informations nécessaires à l'analyse, et les connaissances spécifiques requises ont notamment été mentionnées pour justifier cette position. D'après d'autres participants, l'analyse pourrait plus judicieusement être réalisée par des tiers spécialisés. Enfin, une réponse a souligné que l'analyse n'était jamais pertinente, les modèles d'IA n'étant par nature pas des objets auxquels le RGPD pouvait s'appliquer.

Question 5.2

Les pratiques actuelles vous semblent-elles permettre à l'utilisateur de réaliser l'analyse par lui-même, si cela était exigé de lui, en demandant éventuellement au fournisseur de lui fournir certaines informations ? Dans tous les cas (*open source*, commercialisation « sur étagère », sous-traitance, etc.) ?

À cette question, plusieurs réponses étaient négatives et ont suggéré que le fournisseur réalise l'analyse et en transmette les résultats. Plusieurs justifications pour cette répartition des obligations sont citées :

- la difficulté pour l'utilisateur de distinguer entre les données d'entraînement et les données utilisées pour le RAG ;
- la redondance de l'analyse si plusieurs utilisateurs en ont la charge ;
- la difficulté spécifique au cas open source, où les données sont rarement publiées avec le modèle.

Des mesures spécifiques aux grands modèles de fondation pourraient être exigées de leurs fournisseurs. Certaines réponses ont également soulevé qu'il pourrait être difficile pour le fournisseur également de réaliser ces tests, notamment dans un contexte collaboratif comme pour l'open source, ou lorsque le fournisseur dispose de moyens limités (cas de la recherche).

D'autres réponses ont considéré que l'utilisateur pouvait être en mesure de réaliser l'analyse dans certains cas et lorsque certaines informations lui sont transmises (cf. question 5.1).

L'utilisateur pourrait par ailleurs être en mesure de réaliser certains tests, comme une analyse de la vulnérabilité du modèle (par de l'audit et des tests de *red-teaming*), une étude de la régurgitation, ou une analyse complète lorsqu'il réalise un ajustement du modèle.

Enfin, une réponse a souligné que l'analyse n'était jamais pertinente, les modèles d'IA n'étant par nature pas des objets auxquels le RGPD pouvait s'appliquer.

Question 5.3

Lorsque le modèle est soumis au RGPD, de quels traitements le fournisseur du modèle est-il responsable ? De quels traitements du modèle (tel que son téléchargement, sa manipulation, la mise à disposition, le déploiement, ou l'utilisation) l'utilisateur est-il responsable ? À quelles conditions ?

Plusieurs situations ont été distinguées dans les réponses, mais un accord a semblé émerger sur deux points :

- **le fournisseur a la responsabilité du développement du modèle** (à moins que celui-ci ne lui soit sous-traité par un déployeur) ;
- **le déployeur a la responsabilité du traitement des donnée traitées dans le cadre de l'utilisation du système.**

Concernant la responsabilité du traitement du modèle dans le cas où une mémorisation aurait eu lieu, les réponses ont considéré que la répartition pouvait dépendre des conditions de déploiement du modèle :

- Quand le modèle est mis à disposition en tant que service auprès d'un organisme, le fournisseur est responsable du traitement du modèle.
- Quand le modèle est mis à disposition en tant que service auprès d'utilisateurs finaux pour le compte d'un organisme, la responsabilité pourrait être partagée selon les cas, le déployeur ayant parfois un rôle dans le choix des moyens et des objectifs du traitement.
- Quand le modèle est partagé, le fournisseur pourrait être sous-traitant dans le cas où il serait en charge de l'installation ou de la maintenance**.**
- Quand un ajustement du modèle a lieu, la responsabilité du fournisseur prendrait fin après celui-ci.
- Quand le modèle est publié, le fournisseur devrait être responsable de traitement, puisqu'il serait le seul à pouvoir mettre à jour le modèle et à informer les utilisateurs du modèle de l'exercice d'un droit.
- Quand un contrat existe, les conditions qu'il prévoit devraient prévaloir.
- Quand le modèle est génératif et mis à disposition du grand public par un déployeur, celui-ci devrait avoir une responsabilité similaire à celle des hébergeurs de données.

Quelle que soit la répartition des rôles, il a été suggéré que l'utilisateur soit en charge de signaler les failles observées au fournisseur, et qu'il respecte une durée de conservation du modèle.

D'autres réponses ont toutefois supporté que le traitement du modèle n'impliquait pas de responsabilité au regard du RGPD, que l'utilisateur aurait des difficultés à endosser la responsabilité du traitement du modèle, que l'utilisateur pourrait devenir responsable de traitement uniquement lorsqu'une fuite de données est observée. Il a également été suggéré que le déployeur ne soit responsable que dans le cas où il n'aurait pas respecté les utilisations du modèle prévues dans le contrat ou la licence d'utilisation.

Question 5.4

Le respect des droits relève-il des seuls fournisseurs ou également des utilisateurs du modèle ? Dans quelle mesure ? Quelles techniques permettraient aux utilisateurs de répondre aux demandes d'exercice de droits ? aux fournisseurs ? Quelles techniques

exigeraient en revanche un effort disproportionné ? Quelle coordination entre fournisseur et utilisateur permettrait de garantir la prise en compte des demandes tout au long de la chaîne ?

Tout d'abord, plusieurs difficultés relatives à l'exercice des droits sont identifiées :

- le réentraînement des modèles est jugé disproportionné dans le cas général par plusieurs participants ;
- plusieurs réponses s'accordent sur le manque de maturité des techniques de désapprentissage machine ;
- relier les paramètres du modèle à un individu est difficile et parfois impossible.

Pourtant, il est également constaté que pour appliquer pleinement les droits, le modèle devrait être modifié, et la crainte de l'impossibilité d'exercer ses droits, notamment sur les systèmes génératifs grand public a été soulevée. Le cas de l'*open source* apparaît comme particulièrement difficile, une réponse ayant suggéré que l'ouverture des modèles soit interdite par défaut dans le cas d'une mémorisation, alors qu'une autre réponse propose que le fournisseur en ait la charge et qu'il mette en œuvre une communication avec les utilisateurs pour recevoir les demandes et leur transmettre la réponse apportée.

Ainsi, plusieurs réponses proposent de répartir cette responsabilité entre fournisseur et déployeur, par :

- la transmission des demandes du déployeur au fournisseur, ou à défaut, la transmission du contact du fournisseur par le déployeur à la personne concernée ;
- l'utilisation d'outils ou de mécanismes par le fournisseur permettant de répondre aux demandes adressées au déployeurs ;
- la répercussion par les déployeurs des modifications du modèle effectuées par le fournisseur (ce qui peut entraîner une difficulté dans le cas où le déployeur avait ajusté le modèle du fournisseur), avec le maintien d'un point de contact entre fournisseur et utilisateurs pour cela, y compris dans le cas de l'ouverture du modèle ;
- la démonstration par le fournisseur de l'impossibilité de réidentifier les personnes par un accès au modèle, permettant de déroger à l'application des droits.

Certains cas pourraient toutefois être moins problématiques, comme lorsque le fournisseur est également déployeur et donne accès à son modèle via une API. Dans ce cas, il a été indiqué que le fournisseur pourrait assumer la charge de l'exercice des droits. Dans de nombreuses réponses, il est apparu que cette charge devrait être attribuée au cas par cas, en tenant compte du cas d'usage étudié et de la répartition des responsabilités de traitement. L'un de ces cas particuliers est l'ajustement d'un modèle par un utilisateur : dans cette situation, il est proposé que l'utilisateur soit en charge de répondre aux demandes. Dans les autres cas, le fournisseur semble généralement plus à même de répondre aux demandes, notamment pour les demandes exigeant une action sur les données d'entraînement. Un participant indique que les droits ne peuvent pas être exercés sur les modèles mais seulement sur les données d'entraînement, et liées à l'utilisation.

Enfin, des solutions alternatives sont proposées, comme :

- l'utilisation de filtres sur les entrées et les sorties du système quand le réentraînement n'est pas possible,
- l'anticipation des demandes avant l'entraînement et la mise en œuvre de réponses aux demandes à ce stade.

Plus généralement, un participant a suggéré qu'une discussion sur ce sujet soit initiée par la CNIL avec les fournisseurs des principaux modèles de fondation. Un autre a indiqué que l'élaboration d'une matrice de responsabilités, clarifiant la charge de l'exercice des droits, et plus généralement les obligations du RIA et du RGPD pourrait être utile.

Question 5.5

Dans quelles situations vous semble-t-il disproportionné de ré-entraîner un modèle afin de répondre à une demande d'exercice de droit ? Quels critères ou techniques vous semblent pertinents pour identifier ces situations ?

Aucune des réponses apportées n'a considéré que le réentraînement devait être exigé dans tous les cas. Certaines réponses ont, au contraire, considéré que le réentraînement était disproportionné dans tous les cas, pour des raisons de coût ou parce qu'il ne s'agirait pas d'une solution pertinente pour répondre à une demande d'exercice de droit.

Certains participants ont souligné que l'exercice des droits ne devait porter que sur les données d'entraînement brutes. D'autres ont considéré que cela pouvait être exigé dans les situations les plus à risques ou en cas de manquement lié au traitement de certaines données d'entraînement en raison des coûts financier et environnementaux liés.

Enfin, plusieurs participants ont considéré que cela pouvait être envisagé dans de plus nombreux cas, en tenant compte du risque. Ainsi, devaient être considérés :

- Le niveau de risque pour les personnes lié au traitement, et ainsi l'impact du rejet d'une demande d'exercice de droit (qu'il faut parfois confronter au risque pour les personnes du nouveau traitement de données nécessaire pour le réentraînement).
- L'existence de mesures de prévention permettant de réduire ce risque, comme des filtres sur les entrées et sorties du système.
- D'autres facteurs impactant le coût des réponses aux demandes des personnes concernées, comme la fréquence des requêtes, le coût financier et environnemental du réentraînement, les impacts pour la performance et l'introduction de biais.

Il a également été soulevé que le réentraînement ne devrait jamais être considéré comme disproportionné dans certains cas, comme pour les modèles qui ne sont pas des modèles de fondation, sans garanties supplémentaires (comme la présence de filtres), et lorsque la mémorisation est avérée. En particulier, le réentraînement pourrait être envisagé pour des modèles de petite taille. Au contraire, le réentraînement ne semble pas envisageable dans certains cas, comme en cas d'indisponibilité des données d'entraînement, ou chez les utilisateurs du modèle qui ne sont plus en lien avec le fournisseur ou lorsqu'ils ont réalisé un ajustement du modèle. D'autres réponses ont suggéré qu'une analyse démontrant que les données en question n'ont pas eu d'influence sur l'entraînement, ou la présence de suffisamment de données similaires à celle de la personne concernée dans le jeu d'entraînement permettaient de déroger au réentraînement.

Les alternatives au réentraînement mentionnées sont l'ajustement, le réentraînement régulier (avec un délai qui dérogerait à la durée imposée par le RGPD) avec une suppression immédiate des données brutes de la base d'entraînement.

6. Questions finales

Question 6.1

Au regard des risques de régurgitation et d'extraction de données pour les personnes à partir du modèle entraîné, et en tenant compte des techniques existantes, vous semble-t-il proportionné d'exiger l'analyse du caractère anonyme des modèles d'IA que vous êtes amenés à traiter ? Si oui, quelles techniques d'analyse vous semblent suffisantes ? Si non, quelles contraintes rendent cette analyse disproportionnée ? Merci de justifier en indiquant des éléments de contexte sur votre situation propre.

Un participant a affirmé que cette analyse semblait proportionnée sans considération supplémentaire, un autre a indiqué que l'analyse devait être systématique, mais reposer sur des tests plus ou moins exigeants selon les cas, alors que la plupart des participants ont suggéré une distinction au cas par cas.

Ainsi, un participant a considéré que pour son cas spécifique, cela semblait proportionné au vu du modèle et des données traitées, mais aussi du rôle tenu de fournisseur et déployeur de modèle. Deux autres ont considéré que cela ne devait être exigé que des grands fournisseurs. Un autre participant a considéré que des mesures aménagées pour les fournisseurs de modèles de fondation pouvaient être prévues, alors que le RGPD devrait pleinement s'appliquer sur les modèles ajustés et sur les modèles prédictifs entraînés sur des données

personnelles. Le rôle des fournisseurs semble ainsi particulier puisque l'analyse en amont leur permettait de clarifier les règles pour le reste des acteurs dans la chaîne de l'utilisation du modèle, bien que les utilisateurs puissent réaliser une analyse de risque. Dans le cas où des données sensibles ont été utilisées pour l'entraînement, l'analyse a semblé proportionnée à deux participants, bien qu'il ait souligné qu'elle manque de robustesse et que les solutions préventives plus fiables portant sur les données d'entraînement pouvaient être préférables. Une réponse a souligné que bien qu'elle ne permette pas de répondre de manière binaire à la question de l'application du RGPD à un modèle d'IA, l'analyse est utile puisqu'elle permet d'estimer le niveau de risque du traitement. Il a également été suggéré que l'analyse soit exigée ou non selon des critères contextuels, comme le mode de mise à disposition du modèle, ou les risques connus d'extraction de données.

Lorsque l'analyse a été jugée disproportionnée, plusieurs justifications ont été données :

- la réalisation de tests par l'utilisateur au moyen de requêtes ;
- les moyens mis en œuvre en amont, comme la déduplication des données d'entraînement, l'utilisation de données publiquement accessibles uniquement, et le dimensionnement du modèle au données, suffisent à considérer le risque comme suffisamment faible sans nécessiter d'analyse complémentaire ;
- pour son cas d'usage relatif au ciblage publicitaire, le participant n'a pas connaissance de techniques d'attaques permettant d'extraire des données personnelles des modèles, et il considère que les mesures de minimisation, de pseudonymisation, le réentraînement fréquent des modèles et les restrictions d'accès suffisent à considérer que l'analyse n'est pas nécessaire ;
- les mesures portant sur les données d'entraînement (de déduplication et de pseudonymisation) apportent davantage de garanties, alors que les techniques d'analyse manquent de robustesse ;
- le participant juge que son modèle n'est pas à risque ;
- les modèles d'IA sont par nature des objets auxquels le RGPD ne s'applique pas.

Question 6.2

Avez-vous d'autres réflexions ou remarques dont vous souhaiteriez faire part ?

Les réponses des participants ont permis d'identifier certains sujets demandant des clarifications :

- la répartition des rôles prévus par le règlement IA et le RGPD, notamment pour l'ajustement des modèles ;
- les conditions selon lesquelles un utilisateur pourrait se fier à l'analyse d'un fournisseur indiquant qu'un modèle est anonyme, et plus généralement comment les recommandations de la CNIL permettent à l'ensemble des acteurs de respecter leurs obligations ;
- certaines notions, comme la régurgitation, l'inférence, la mémorisation, si un modèle peut être qualifié d'anonyme, ou de pseudonyme, ou encore la qualification des sorties des modèles portant sur un individu mais non liées à la mémorisation, ;
- les techniques de sécurité permettant de protéger les modèles ;
- l'application des recommandations de la CNIL aux cas spécifiques de l'apprentissage en continu, du *retrieval augmented generation* (RAG) et des systèmes interconnectés.

D'autres ont cherché à préciser ou appuyer certains points spécifiques :

- Il a été pointé que les publications mentionnées par la CNIL afin de démontrer l'existence d'un risque d'extraction de données personnelles d'un modèle entraîné étaient critiquables. Ces études ne démontreraient pas l'existence d'un risque en pratique car les attaques développées sont mises en œuvre dans des conditions très spécifiques (avec un accès à certaines données d'entraînement), par des personnes aux compétences techniques et aux moyens importants, et fournissent des résultats restreints (en volume de données extraites) et dont la fiabilité n'est vérifiée que grâce à l'accès aux données d'origine.
- La recherche sur les sujets abordés doit être poursuivie, notamment via un état des lieux des connaissances scientifiques.

- Davantage de coopération entre les fournisseurs et utilisateurs devrait être mise en place.
- Les développeurs de modèles et utilisateurs devraient être davantage formés et sensibilisés à la sécurité et à la confidentialité.
- Les recommandations de la CNIL auront un impact sur l'écosystème, et notamment sur l'innovation, qu'il est nécessaire de prendre en considération. Une attention devrait ainsi être portée à l'accessibilité des outils (financière et en termes de prérequis) et aux données, notamment pour la recherche.
- Les risques pour les modèles génératifs et prédictifs étant sensiblement différents, une distinction devrait être effectuée.
- Le risque d'extraction de données personnelles n'est pas celui dont les conséquences sont les plus importantes pour les personnes et pour la société, notamment en comparaison des risques liés à la cybercriminalité, aux hypertrucages, à l'hameçonnage, à la reconnaissance faciale ou aux biais. Le risque lié à l'utilisation des grands modèles de langage pour inférer des informations sur une personne (sans mémorisation, comme documenté dans Staab et al. 2024) font également l'objet de préoccupations.
- Le raisonnement selon lequel les modèles d'IA contiendraient des données personnelles n'est pas le bon, et ainsi il est estimé que les positions prises par l'autorité de protection de données régionale de Hambourg et celle du Danemark sont proportionnées.
- Considérer que les modèles d'IA sont en dehors du périmètre d'application du RGPD n'entraînerait pas de faille dans l'application du RGPD aux traitements d'IA puisque les phases d'entraînement et d'utilisation y seraient toujours soumises.
- Le développement de mesures au niveau applicatif doit être privilégié, comme les politiques d'usage, les interactions du système avec les données personnelles, le tatouage de données, l'ajustement des modèles, les filtres et classifieurs utilisés sur les réponses, le contrôle des utilisateurs sur le système, et la cybersécurité.