Octobre 2024

Résumé de la contribution du Laboratoire d'Intelligence Artificielle de Criteo à la consultation de la CNIL sur l'application du RGPD au développement des systèmes d'intelligence artificielle et aux modèles d'intelligence artificielle

A propos du laboratoire d'IA de Criteo

Criteo est le leader européen de la publicité digitale basée sur l'intelligence artificielle (IA). Avec plus de 100 ingénieurs et chercheurs, le laboratoire d'IA de Criteo est l'un des pionniers français de l'innovation en machine learning et l'un des plus grands laboratoires de recherche et développement privé européen en termes de publications.

I. Développement des systèmes d'intelligence artificielle (IA)

1. Utilisation de la base légale de l'intérêt légitime pour développer un système d'IA

- Les travaux de **recherche scientifique** tout autant que les **activités de nature commerciale** doivent pouvoir mobiliser la base légale de l'intérêt légitime. Les projets privés alimentent la recherche scientifique en IA tout en la finançant, notamment par le biais de thèses de doctorat. Ces projets de recherche sont ensuite déclinés dans des cas d'usages concrets et bénéficient le plus souvent à un écosystème de partenaires, clients ou utilisateurs privés ou publics voire à la société dans son ensemble. La séparation "intérêt commercial/intérêt public" n'est pas pertinente pour déterminer le niveau de risque d'un projet ou son intérêt légitime.

- Les **bénéfices du système d'IA** doivent être pris en compte dans l'évaluation de l'intérêt légitime, notamment si le système d'IA permet l'accès à l'information ou la liberté d'expression. Mais cette base légale ne doit pas augmenter les **contraintes en matière d'IA**, telle que par exemple la mise en place d'un opt-out discrétionnaire.

2. Spécificité des modèles en open source

- Les modèles en open source jouent un rôle important pour l'innovation en Europe en permettant aux **entreprises européennes de développer leur(s) propre(s) modèle(s)**. L'open-source favorise par ailleurs le contrôle et l'amélioration du modèle par les pairs et donc la transparence, la réduction des biais et la sécurité des modèles. Les modèles open source présentent aussi des avantages pour la communauté notamment en termes de formation. Afin de favoriser l'innovation, il est important que les **licences de diffusion des modèles open-source autorisent la réutilisation commerciale** des modèles.

- La notion d'open source doit reposer sur **des critères objectifs et l'état des développements technologiques** tout en étant aligné avec le règlement sur l'IA (RIA), Ces critères sont la diffusion du modèle et de ses poids, accompagnée d'une documentation et d'une analyse du modèle (ce qui exclut les modèles uniquement accessibles par API). L'ouverture de la procédure d'entraînement du modèle (diffusion du code d'entraînement et de sa documentation) ou l'ouverture du jeu de données doivent rester optionnels, car permettant aux fournisseurs de ces modèles de rester compétitifs.

- L'ouverture du jeu de données d'entrainement n'est pas une condition suffisante pour **évaluer un modèle.** En ce qui concerne la détection des biais par exemple, celle-ci nécessite principalement l'observation des résultats retournés par le modèle. Mais avant toute chose, il convient d'auditer la présence de biais dans les données avant entraînement.

- Les obligations relatives à l'information des personnes et l'exercice des droits doivent prendre en compte la **spécificité des modèles en open-source** afin de ne pas favoriser les modèles propriétaires : ne pas nécessiter une fermeture du modèle et un contrôle sur les utilisations, ce qui est contraire à la nature même de l'open-source. Des **protocoles ouverts et collaboratifs et la standardisation** pourraient permettre à tous les acteurs de répondre aux demandes.

- Les modèles ouverts **favorisent la concurrence** en donnant des moyens aux plus petits acteurs. Dès lors, toute obligation supplémentaire à respecter de la part des diffuseurs des modèles open source pourrait avoir un impact négatif sur la concurrence et la recherche scientifique. Il faudrait au contraire que les régulateurs **accompagnent les contributeurs open source** pour les aider à mettre œuvre la réglementation existante et prévoir davantage d'exceptions aux obligations pour la recherche scientifique.

3. Information et exercice des droits des personnes

- Les obligations relatives à l'information et l'exercice des droits des personnes doivent être mises en balance avec la **protection des secrets d'affaires** du développeur du système.

- La fourniture d'information et l'exercice des droits s'applique **aux données personnelles utilisées pour entrainer le modèle et non au modèle** en tant que tel (lequel peut être entrainé avec des données non-personnelles et donc non soumis au RGPD.

- L'information et l'exercice des droits des personnes doit être adapté à la **spécificité de chaque modèle**. Dans ce cadre, s'il est nécessaire que les personnes reçoivent une information simple, il est aussi important que les pouvoirs publics et les acteurs de l'IA mettent en place des campagnes de sensibilisation et de vulgarisation sur le fonctionnement des systèmes d'IA.

- L'exercice des droits de suppression, d'opposition ou le retrait du consentement peuvent créer des tensions avec **la représentativité des données** utilisées pour entrainer un modèle et entrainer un biais du modèle ou des résultats inexacts. L'exercice du droit de rectification peut être problématique dans le cas où les données ont été modifiées par du bruit pour mieux les protéger.

- Toute contrainte relative au **réentrainement du modèle lié à l'exercice des droits** des personnes doit respecter le principe de proportionnalité en ce qui concerne notamment les coûts engendrés, la faisabilité technique en fonction des caractéristiques du modèle, ou encore la mobilisation excessive de ressources informatiques et énergétiques. Il se peut par ailleurs que les données initiales ne soient plus accessibles, ou que le ré-entraînement du modèle ne permette plus d'obtenir des performances équivalentes. La répercussion de la modification du modèle par ses utilisateurs peut aussi être difficile. Dans ces cas, en l'absence de risque élevé pour la personne, il serait déraisonnable de demander un réentraînement du modèle, en attendant l'**émergence d'une technique de désapprentissage machine robuste et peu coûteuse**.

4. Annotation des données

- L'application du RGPD ne doit pas aboutir à favoriser l'**apprentissage non supervisé par rapport à l'apprentissage supervisé**. Si l'annotation des données a généralement peu d'impact sur les personnes, les obligations du RGPD devraient être modulées en conséquence. Il est difficile par exemple de voir comment le droit à la portabilité aura vocation à s'appliquer dans ce contexte.

II. Développement des modèles d'intelligence artificielle

1. Typologie des risques de réidentification

- La réidentification peut résulter d'une **utilisation normale** du modèle, en particulier d'IA générative. La réidentification peut aussi résulter d'une **volonté de réidentification soit à des fins légitimes d'audit, soit dans un but malveillant**. Elle dépend du contexte d'accès au modèle et aux données qui peut aller de l'accès à l'intégralité du modèle (white box) ou aux seules caractéristiques d'entrée (black-box), ce qui rend l'attaque plus incertaine.

2. Mémorisation lors de l'entraînement du modèle

- La mémorisation est une composante importante pour la bonne performance du modèle. Le concept en tant que tel est toujours **débattu au sein de la communauté scientifique**. Il peut prêter à confusion car la mémorisation ne consiste pas en la mémorisation de la donnée en tant que telle, mais en **l'observation de l'influence d'un point de donnée** sur la performance du modèle[1] selon s'il est inclus on non dans le jeu de données d'entraînement.

- L'**influence de la mémorisation sur les attaques de confidentialité n'est pas binaire** : il peut y avoir une corrélation entre les points de données "mémorisés" par un modèle et ceux étant plus sujet à des attaques de confidentialité. Mais la mémorisation ne se définit que par la valeur d'une mesure pour un point de donnée relativement aux autres alors que le taux de réussite d'une attaque de confidentialité est plus simple à mesurer.

- Quel que soit le type de modèle, **certains points de données sont plus susceptibles d'être mémorisés**, notamment lorsque la donnée est rare, suit une loi de probabilité à queue lourde[2] et qu'elle est répétée dans le jeu de données. Toutefois, un modèle génératif est par nature plus susceptible de mémorisation et plus simple à attaquer qu'un modèle discriminatif[3], du fait de la complexité de ses sorties.

3. Vraisemblance d'une régurgitation ou de l'extraction de données

- Les **cas de régurgitation spontanée** de données pour les IA génératives sont rares. Sans accès à la donnée (ou à sa distribution) d'origine, il est difficile de les distinguer d'autres sorties possibles des modèles.

- Les **cas d'extraction de données** quel que soit le type d'IA ne peuvent résulter que d'attaques ciblées malveillantes qui sont généralement coûteuses à mettre en œuvre. La connaissance par l'attaquant du modèle est un facteur important de réussite de l'attaque. Bien qu'il semblerait plus facile d'attaquer un modèle ouvert que fermé, les modèles ouverts sont plus faciles à auditer et donc potentiellement mieux sécurisés.

- Les cas de régurgitation ou d'extraction de données doivent être clairement distingués des cas de fuites de données. Ces cas sont **extrêmement rares et présentent de très faibles risques**, d'autant plus si les données ont été pseudonymisées. De plus, leur nature probabiliste rend incertaine la confirmation des éventuelles informations extraites par l'attaquant à moins de déjà connaître la donnée. Les auteurs de ces attaques peuvent également être sanctionnés pénalement.

---

[1] Par exemple en comparant la performance du modèle avec et sans ce point de donnée.
[2] C'est-à-dire, lorsque les données peuvent prendre plus fréquemment des valeurs anormales voire inattendues.
[3] Un modèle génératif est conçu pour modéliser et apprendre, à partir d'un jeu de données, des frontières entre différentes classes ou catégories.

- En dehors certaines applications à haut risque telles que définies par le RIA, la connaissance de la présence de la donnée dans le jeu d'apprentissage a généralement peu de conséquences. Dans le secteur de la **publicité en ligne**, connaître la présence d'un utilisateur dans la base de données d'entraînement n'apporte pas à l'attaquant d'information supplémentaire par rapport à la donnée elle-même, si ce n'est que l'utilisateur a consenti à la collecte de ses données ou qu'il lui a été présenté une publicité.

4. Critères d'analyse du risque de régurgitation ou d'extraction

- Le **nombre de paramètres** du modèle au regard du volume de données ne semble pas être un critère adéquat d'analyse du risque car (1) le poids mémoire des données ou leur nombre ne permet pas de mesurer la quantité d'information et (2) ce critère dépend du modèle utilisé et est insuffisamment relié aux risques de régurgitation et d'extraction. La "sur-paramétrisation" du modèle est essentielle à la bonne performance de l'apprentissage profond, lequel ne présente pas plus de risques pour les données que d'autres techniques d'apprentissage. Par la complexité abstraite de son modèle, il serait même plus avantageux que d'autres techniques plus explicites comme la recherche de plus proches voisins. Enfin, le surapprentissage ne peut pas être assimilé à de la mémorisation, même si les deux notions peuvent se recouper.

- La liste des critères d'analyse du risque ne doit pas aboutir en pratique à exiger une analyse du risque de régurgitation pour tous les modèles qui sont concernés en application réelle par certains de ces critères. Il conviendrait plutôt **d'identifier les cas où les risques de régurgitation et d'extraction sont faibles**. Au regard de la rareté des cas de régurgitation et d'extraction en pratique, très peu de modèles devraient être soumis à une analyse de risques poussée.

5. Techniques permettant d'analyser les risques de régurgitation et d'extraction

- Le simple risque de régurgitation ou d'extraction ne devrait pas à lui seul déclencher la qualification de donnée personnelle. Une telle approche aurait pour effet de créer une **insécurité juridique** pour les acteurs dont le régime juridique dépendrait d'un risque externe qu'ils ne maitrisent pas, mais généralement très faible et dépendant d'acteurs malveillants enfreignant la loi et dotés de moyens importants.

- Les modèles résultant d'apprentissage machine **ne contiennent pas directement d'identifiant attribuable à un point de données ou un individu** dans leurs paramètres. L'influence des points d'entraînement est agrégée dans leurs paramètres, qui sont des représentations mathématiques non interprétables. Si les techniques existantes ne permettent pas de déceler un risque significatif de régurgitation ou d'extraction, alors le modèle devrait **être considéré comme anonyme jusqu'à preuve du contraire.** Il faudrait de façon générale éviter une approche fondée sur le **risque zéro** sinon aucun des modèles visés ne serait considéré anonyme.

- La simple détention d'un modèle ayant « mémorisé » des données personnelles ne devrait **pas être considérée comme une conservation de données personnelles** du fait de la nature agrégée, abstraite, non interprétable – et donc non indentifiable – des poids du modèle. Des exceptions pourraient être envisagés en cas d'un risque de réidentification élevé à coût faible, si le modèle est à haut risque tel que défini par le RIA ou lorsque sa sortie est à caractère identifiant.

6. Responsabilité des acteurs

- L'obligation d'analyser le caractère anonyme du modèle viendrait s'ajouter aux obligations existantes tout en créant un **régime plus contraignant pour l'IA** alors que le RGPD est neutre technologiquement. Cette analyse ne pourrait être effectuée que par le fournisseur du modèle. L'utilisateur n'a pas accès aux moyens nécessaires pour conduire cette analyse (accès direct au modèle, aux données ou au code d'entraînement, moyens de calcul). En cas de fine-tuning du

modèle sur de nouvelles données personnelles, l'analyse ne devrait pas porter sur les données d'entraînement initiales.

- **L'open-source et la recherche** sont des cas particuliers car ils pourraient manquer de moyens pour réaliser l'analyse du caractère anonyme du modèle. Cette obligation pourrait donc favoriser les modèles propriétaires à accès restreint.

- **Le respect des droits des personnes** relève davantage de la responsabilité du fournisseur que de l'utilisateur dont l'obligation devrait se limiter à renvoyer la personne vers le fournisseur. Si l'utilisateur opère lui-même le modèle, il devrait répercuter les modifications sur son modèle local, à la manière d'un correctif de sécurité appliqué à un logiciel. Cette répercussion est compliquée si l'utilisateur a « fine-tuné » le modèle sur d'autres données.

7. Analyse au regard des modèles entrainés par Criteo

- Les modèles entraînés par Criteo effectuent principalement des **tâches de prédiction et de recommandation publicitaires en ligne**. Il ne semble pas nécessaire d'exiger l'analyse de leur caractère anonyme :
    - Les données personnelles utilisées dans la phase d'entraînement des modèles sont sélectionnées selon le principe de minimisation via du *negative sampling* ;
    - Les données sont complètement pseudonymisées et hashées empêchant toute réidentification directe ;
    - La durée de vie des modèles est courte, ceux-ci étant réentraînés régulièrement sur de nouvelles données ;
    - Les modèles ne sont pas accessibles hors de l'entreprise ;
    - Les résultats auxquels Criteo a accès sont le résultat de l'enchère à laquelle Criteo a participé (perdu/gagné) et le produit/service recommandé ;
    - Tout attaque de confidentialité nécessiterait une connaissance approfondie de notre technologie, ce qui soulèverait des enjeux liés au secret industriel ;
    - Il n'y pas de manière accessible d'extraire des données d'entraînement de nos modèles ;
    - Nos règles internes interdisent ces pratiques ;
    - Nos modèles ne sont pas génératifs, excluant les risques de régurgitation et d'extraction.

8. Remarques finales

- Les questions posées par la CNIL concernent des problématiques scientifiques récentes et non résolues. Les réponses à ces questions devraient être précédées d'une analyse de la littérature et d'un **état des lieux circonstancié et consensuel** des connaissances scientifiques par le biais de **comités représentatifs de la communauté scientifique** française et européenne dont les débats seraient ouverts, contradictoires et transparents.

- L'application du RGPD à l'IA doit être basée sur une évaluation des risques raisonnables tenant compte de l'impact du système d'IA dans son contexte, à l'exclusion de la mise en œuvre d'un **principe de précaution qui pourrait freiner l'innovation en IA**. Le risque que des modèles soient exclus du marché européen du fait de l'incertitude juridique résultant de l'interprétation du RGPD doit également être pris en compte.

- La présence potentielle de données personnelles pourrait être gérée par l'utilisation de **techniques de pseudonymisation** qui permettent de concilier la protection des données et la conservation de leur utilité. Il est important que leur développement soit encouragé.

***