

Chers membres du réseau de l'ancienne FSFA, chers intéressés,

C'est avec plaisir que je me souviens de l'intéressant [congrès du 21.3.2024](#). Un grand merci à tous ceux qui y ont participé, en ligne ou sur place. J'ai été particulièrement ravi des discussions que j'ai pu avoir avec des membres du réseau de l'ancienne FSFA en marge du congrès à Zurich.

Dans les prochains courriels, je continuerai à éclairer des aspects intéressants de l'IA.

Les modèles linguistiques de l'IA se dégradent-ils à cause des données générées par l'IA ?

L'Internet se remplit de plus en plus de textes générés par l'IA. Qu'est-ce que cela signifie pour l'entraînement des modèles linguistiques de l'IA ? Deux publications traitant de cette question arrivent à la conclusion suivante : si un dataset d'entraînement d'un modèle linguistique de l'IA contient trop de textes générés par l'IA, les résultats se dégradent au fur et à mesure que l'entraînement se répète. Le titre de la première publication est révélateur : [The Curse of Recursion: Training on Generated Data Makes Models Forget](#). (La malédiction de la récursion : l'entraînement sur des données générées fait oublier les modèles ou les rend inopérants). Ce document a été rédigé par une équipe d'auteurs issus principalement d'universités britanniques. Une deuxième publication menée en Angleterre, qui a entraîné des modèles avec des images, arrive à une conclusion similaire. Les résultats ont été présentés sous le titre : [Towards Understanding the Interplay of Generative Artificial Intelligence and the Internet](#).

Est-ce vraiment possible ?

Il est difficile de distinguer les textes ou les images générés par l'IA de ceux créés par l'homme. Le problème réside précisément dans le fait qu'il n'existe pas encore de méthode sûre, même avec une sorte de « filigrane », pour distinguer les textes générés par l'IA des autres. Que se passe-t-il ici – y a-t-il une sorte de boucle de rétroaction ? [Ben Lutkevich](#), explique [ici](#) (avec vidéo explicative, 1', aussi directement sur [YouTube](#)) le phénomène sur [WhatsIs?](#), un site de l'entreprise américaine TechTarget. Lutkevich donne à la fin le conseil de recourir provisoirement pour l'entraînement à des données antérieures à 2023, avant que l'Internet ne soit « pollué » par des données générées par l'IA (Data pollution).

Les géants de la technologie s'assurent des données de haute qualité créées par l'homme

Les géants de la technologie sont – sous pression – apparemment prêts à payer pour l'utilisation des données. OpenAI, comme d'autres géants de la technologie, avait dans un premier temps utilisé sans autorisation les données disponibles sur Internet pour l'entraînement de leurs modèles linguistiques de l'IA. Se posent alors les questions du droit d'auteur et d'une indemnisation adéquate pour l'utilisation des données. Les données de grande valeur des entreprises de médias présentent un intérêt particulier. Déjà en décembre 2023, OpenAI s'est exprimé à la [Chambre des Lords](#) britannique (4^e point) en disant qu'il était impossible d'entraîner des modèles d'IA sans matériel protégé par le droit d'auteur, raison pour laquelle OpenAI a conclu des partenariats avec des maisons de médias qui sont avantageux pour les deux parties. En effet, dès l'été 2023, OpenAI et l'agence de presse américaine, l'[Associated Press \(AP\)](#), ont conclu un partenariat qui permettait à OpenAI d'accéder aux archives de l'AP jusqu'en 1985. En décembre 2023, le groupe de presse allemand [Axel-Springer](#) a conclu un partenariat avec OpenAI. Le montant payé par OpenAI n'a pas été publié.

Apparemment, le New York Times et OpenAI ne sont pas parvenus à un accord. Comme l'a annoncé l'agence de presse britannique [Reuters](#) le 27.2.2024, la maison de médias a déposé une plainte. Depuis, les deux sont en conflit juridique. Une action en justice couronnée de succès devrait encourager d'autres entreprises de médias plus petites à suivre l'exemple. Le fait est que les applications d'IA représentent un énorme business pour les géants de la technologie. Selon le [New York Times](#) du 16.2.2024, la valeur de l'entreprise OpenAI a triplé en dix mois pour atteindre 80 milliards de dollars.

D'autres géants de la technologie dans la tourmente

Mais d'autres géants de la technologie sont également sous pression. Selon le [journal télévisé de l'ARD](#) du 28 février 2024, 32 entreprises de médias de 17 pays européens ont poursuivi Google en justice pour obtenir 2,3 milliards d'euros de compensation pour les pertes de revenus publicitaires. En 2022, Google a réalisé un chiffre d'affaires de 224,5 milliards avec la publicité en ligne.

Avec nos salutations les meilleures,
Pour le réseau de l'ancienne FSFA : Hanna Muralt Müller

Nouveau droit de la protection des données : Si vous ne souhaitez plus recevoir cet e-mail, veuillez me contacter !