LesEchos.fr

www.lesechos.fr
Pays : France
Dynamisme : 0

Visualiser l'article

Le mariage explosif de nos données et de l'IA

Plusieurs scandales sur l'utilisation de bases de données ont ponctué l'actualité ces derniers temps. L'utilisation de données personnelles pour entraîner des intelligences artificielles sans le consentement des utilisateurs fait de plus en plus débat.

L'année dernière, le règlement européen sur la protection des données (RGPD) établissait un nouveau cadre pour le traitement automatique des données.

Comment savoir à quoi nos données, quand elles sont utilisées pour entraîner des intelligences artificielles (IA), peuvent réellement servir ? A repérer des tchats dans des vidéos YouTube ? A nourrir les algorithmes de l'armée américaine ? A faciliter l'oppression de la minorité ouïgoure en Chine ? Impossible de répondre, mais le sujet devient de plus en plus sensible.

Si l'IA fait autant parler d'elle ces dernières années, c'est principalement grâce à l'apprentissage automatique. C'est-à-dire la capacité pour l'ordinateur à apprendre tout seul ou presque : on lui fournit une grande masse de données, des portraits avec un nom par exemple, et à force d'entraînement, il va être capable d'identifier des personnes. Pendant longtemps, ces fichiers ont été rassemblés par les chercheurs dans des jeux de données (datasets) servant au départ au monde académique et à la R&D.

Mais les données, qui étaient jadis utilisées pour un traitement inoffensif (l'exemple des tchats dans les vidéos est véridique), peuvent aujourd'hui servir à des usages bien plus problématiques : parfois, nous figurons dans ces jeux de données sans même le savoir.

La question n'est pas que mon image soit utilisée, c'est comment elle l'est.

C'est ainsi que Jillian York, activiste au sein de l'ONG Electronic Frontier Foundation, a découvert des photos d'elle, prises dans des lieux privés, au sein d'une base de données. *« Le problème n'est pas que mon image soit utilisée, c'est la façon dont elle elle l'est »* , indiquait-elle dans une enquête du « Financial Times » .

Le jeu de données en question a été mis au point par l'Iarpa (Intelligence Advanced Research Projects Activity), une agence gouvernementale américaine développant des projets pour le monde du renseignement. Pour constituer cette base, les scientifiques de l'Iarpa ont fait au plus simple : ils sont allés chercher sur le Web des images sous licence Creative Commons, ce qui les rend quasiment libres de droits. Sauf que les personnes prises en photo n'ont jamais été informées.

RGPD : application perfectible

Le règlement européen sur la protection des données (RGPD), entré en vigueur l'année dernière , est censé nous préserver de ces usages. L'article 9 dit clairement qu'il faut que *« la personne concernée* [ait] *donné son consentement explicite au traitement »* . Il est aussi nécessaire de définir la « finalité » et la « durée » d'utilisation.

Au début du mois, Microsoft a supprimé une base de données de 10 millions de photos utilisées pour entraîner des systèmes de reconnaissance faciale, MS-Celeb. Celle-ci avait notamment été utilisée par

Safran pour des recherches sur l'identification de personnes, ou par l'Université nationale des technologies de défense en Chine.

Ce dataset contenait les portraits de plus de 100.000 individus. *«* [Microsoft l'a] *probablement enlevé parce que* [ses] *avocats ont exprimé des inquiétudes de ne pas avoir de base juridique pour traiter ce type de données au regard de l'article 9 du RGPD »* , a indiqué au « Financial Times » Michael Veale, chercheur au Alan Turing Institute.

Lire aussi :
Intelligence artificielle : la crise de confiance

« Une entreprise de l'Union qui fait usage de données sans avoir prévenu les individus, même non européens, risque de se mettre en délicatesse avec le RGPD , ajoute l'avocate Jeanne Bossi-Malafosse, du cabinet Delsol. *Il en va de même de toute société, même extracommunautaire, qui utiliserait des données de résidents européens sans les avoir informés de la finalité. »*

La loi devrait donc fournir une base suffisante pour nous prévenir de l'utilisation présente ou future de nos informations personnelles ; à nous de faire le tri… encore une fois en théorie. Force est de constater que c'est loin d'être évident. Et c'est sûrement le problème sur lequel devrait se pencher une saison 2 du RGPD : le traitement algorithmique.

Mannequin entraîneur d'IA

L'opinion internationale exprime de plus en plus ses inquiétudes : en janvier dernier, la consultante Kate O'Neill postait, d'un tweet devenu viral à propos du « 10 Year Challenge », un défi entre amis consistant à publier sur Facebook ou Instagram deux photos de soi à dix ans d'intervalle : *« Je me demande comment toutes ces données pourraient être analysées pour entraîner des algorithmes de reconnaissance faciale et d'estimation de l'âge »* , lançait-elle *« en plaisantant à moitié »* .

Un autre challenge a d'ores et déjà permis d'améliorer des systèmes d'IA : le « Mannequin Challenge » , qui consistait à filmer un groupe de personnes immobiles en se déplaçant autour d'elles. Le 23 mai, Google AI a publié sur son blog les résultats d'un nouvel algorithme de prédiction de la profondeur sur des images en deux dimensions (ce qui permettrait de les passer en 3D). Pour entraîner le modèle, les chercheurs ont utilisé environ 2.000 vidéos du Mannequin Challenge sur YouTube.

Lire aussi :
Les 10 recommandations de l'OCDE pour l'intelligence artificielle

Ce qui pose la question d'une pratique répandue en informatique : le fait de « crawler » le Web, c'est-à-dire d'utiliser des logiciels pour récupérer automatiquement des données, comme ce fut le cas pour constituer MS-Celeb ou le dataset du Mannequin Challenge. Le RGPD prévoit toutefois que le consentement n'est paradoxalement pas requis quand *« le traitement porte sur des données à caractère personnel qui sont manifestement rendues publiques par la personne concernée »* .

Un tiers peut-il donc réutiliser nos photos Facebook publiques ? *« Je pense que non,* indique Jeanne Bossi-Malafosse. *Car c'est dans ce cas une réutilisation, et si la question du consentement peut se poser, la personne doit être informée. »* Un texte, deux interprétations.

www.lesechos.fr
Pays : France
Dynamisme : 0

Visualiser l'article

Le site Megapixel , de l'activiste américain Adam Harvey, permet de voir en un coup d'oeil l'usage de quelques bases de données (celles qui sont citées dans des articles scientifiques). Bientôt, il envisage de permettre à tout un chacun de taper son nom dans une barre de recherche, pour voir dans quel dataset il apparaît, quelles données ont été récupérées, et à quoi elles ont servi. Les internautes risquent d'avoir quelques surprises.