Deepfake : la Darpa passe à la contre-offensive

Les manipulations de l'information dopées à l'intelligence artificielle (IA) se multiplient dans des vidéos et enregistrements audio circulant sur les réseaux sociaux.

L'humain et la machine peuvent-ils distinguer ces fausses informations alimentées par une IA ? C'est ce que le Département américain de la défense tente de déterminer dans le cadre de recherches financées par la Darpa (Defense Advanced Research Projects Agency), rapporte la MIT Technology Review.

Ce été, des spécialistes en « forensique » et numérique se réuniront pour un concours. Ils devront réaliser des vidéos, des images et des sons modifiés à l'aide d'une IA. Chaque équipe tentera d'identifier les anomalies dans les projets proposés par l'équipe adverse. Le contenu le plus réaliste remportera la mise.

Le challenge sera ainsi l'occasion pour les participants d'étudier des « deepfakes », ces vidéos dans lesquelles le visage d'une personne est déplacé sur le corps d'une autre. Les mouvements et les paroles étant modifiés pour rendre l'ensemble le plus réaliste possible.

Le procédé, popularisé sur Reddit, permet donc de détourner l'image d'individus. Sans surprise, la technologie a déjà été utilisée pour produire des vidéos à caractère pornographique. Elle peut l'être aussi pour manipuler les propos et l'action de personnalités politiques, entre autres.

Réseaux antagonistes génératifs (GANs)

Dans ce contexte, les chercheurs de la Darpa s'intéressent plus particulièrement aux réseaux antagonistes génératifs (GANs, generative adversarial networks). Ces algorithmes d'apprentissage non-supervisé qui permettent de générer des images avec un niveau élevé de réalisme.

Des images dont les anomalies sont très difficiles à repérer automatiquement, selon la Darpa. « En théorie, si vous opposez à un GAN toutes les techniques que nous connaissons pour le détecter, il peut passer toutes ces techniques », a expliqué David Gunning, responsable du programme Media Forensics (MediFor) de l'agence fédérale américaine. « Nous ne savons pas s'il y a une limite. C'est flou », a-t-il ajouté.

La Darpa indique également sur la page du programme MediFor que les outils utilisés aujourd'hui manquent de robustesse. Ils ne sont pas assez évolutifs et couvrent uniquement certains aspects de l'authentification de médias, selon elle.

« Mais une plateforme de bout en bout permettant d'effectuer une analyse légale complète et automatisée n'existe pas », a insisté l'agence américaine de recherche avancée des projets de défense.

(crédit photo © by Donald Tong from Pexels)