Comment repérer et comprendre les contenus générés par l'IA ou modifiés numériquement ?

Des modèles d'IA accessibles et faciles à utiliser peuvent aider les gens à apprendre et à créer du contenu, mais ils peuvent aussi amplifier les risques de désinformation et de mésinformation pour les sociétés ouvertes et le discours démocratique. Il est important d'éviter que nos espaces d'information partagés ne soient encombrés par des informations erronées générées par l'IA et altérées numériquement.

Une partie de la solution réside dans les nouvelles technologies, telles que les initiatives de provenance des contenus et les logiciels de détection. Mais les solutions technologiques sont loin d'être parfaites et nous avons besoin du travail de vérificateurs de faits indépendants pour fournir à la société un ensemble partagé de faits vérifiés.

Voici un bref aperçu de ce que font les vérificateurs de faits professionnels et indépendants pour identifier et démystifier les fausses informations et la désinformation générées par l'IA, et de ce que vous pouvez apprendre d'eux.

Les contenus générés par l'IA se multiplient

Aujourd'hui, les fausses informations et la désinformation générées par l'IA ne représentent qu'un faible pourcentage de toutes les affirmations examinées par les vérificateurs de faits professionnels et indépendants. Les contenus édités numériquement sont plus fréquents dans le travail des vérificateurs de faits.

Mais : dans une enquête interne des membres d'EFCSN, la plupart des vérificateurs de faits ont convenu que le contenu généré par l'IA et modifié numériquement ne fera que gagner en importance à l'avenir. Des exemples récents dans le contexte des élections européennes confirment ce pronostic.

COMPRENDRE : L'altération numérique désigne toute forme de contenu qui a été modifié de manière significative pour manipuler ou changer le message qu'il véhiculait à l'origine, y compris les modifications apportées par les outils d'intelligence artificielle. Cela n'inclut pas les modifications visant à améliorer la clarté ou la qualité.

L'expression "généré par l'IA" désigne toute forme de contenu créé par un système d'intelligence artificielle.

Informations fournies par **AI@EUElections**, financé par Meta et géré par le **European Fact-Checking Standards Network** en association avec ses membres et des experts externes.

La technologie évolue rapidement, mais nous ne pouvons pas compter uniquement sur elle.

Les experts en IA et les vérificateurs de faits professionnels sont d'accord : Les outils de détection de l'IA ne suffisent pas à détecter ou à démystifier les contenus générés par l'IA ou modifiés numériquement.

Avant d'utiliser un détecteur, les experts recommandent de se familiariser avec les générateurs et les détecteurs de contenu d'IA. En comprenant comment les modèles sont formés et en ayant quelques notions de statistiques, les vérificateurs de faits peuvent commencer à reconnaître les forces et les faiblesses d'un outil et ses chances de succès. **Néanmoins, les outils peuvent constituer un point de départ utile.**

Provenance du contenu
Des initiatives telles que les spécifications C2PA peuvent contribuer à attester la source et l'historique du contenu médiatique, mais le filigrane et la vérification ne sont pas irréprochables.

Comment la désinformation par l'IA affecte-t-elle les gens ?

"Chaque fois que vous réagissez, si je puis dire, avec vos tripes, cela court-circuite votre réflexion".
– Christine Dugoin*

PSYCHOLOGIE : Les opérations d'influence sont souvent conçues pour tirer parti des préjugés psychologiques.

Comprendre ses propres préjugés et ceux de son public peut aider à contrer la désinformation.

OBJECTIFS : Pourquoi un acteur malveillant pourrait-il s'appuyer sur l'IA pour créer ou diffuser de la désinformation ? Quel est l'impact recherché dans le monde réel ?

- ➔ Étendre leur portée à un autre pays ou à une autre communauté ?
- ➔ Éviter d'être détectés ou submerger les vérificateurs de faits en générant de nombreuses variantes d'affirmations similaires ?
- ➔ Influencer les pensées ou les croyances en établissant leur crédibilité par le biais de réseaux de comptes non authentiques ?

** Christine Dugoin est chercheuse en influence informationnelle à La Sorbonne.*

Le démenti nécessite une approche à multiples facettes et une compréhension nuancée

Si les outils de détection ne fonctionnent pas, qu'est-ce qui fonctionne ? Il est important de comprendre le contexte d'une affirmation autant que son contenu. Les vérificateurs de faits professionnels sont des experts qui maîtrisent les techniques d'investigation nécessaires. Voici quelques conseils.

> *"Les outils de détection ne fonctionneront jamais à 100 % - je ne pense pas que ce sera le cas un jour".*
> - Henk van Ess**

CONSIDÉRER LA SOURCE : Pouvez-vous confirmer son identité ? De quoi parle-t-elle et que partage-t-elle ? Qui interagit avec son contenu ? Quel effet ce contenu peut-il avoir sur les lecteurs ?

ÉTABLIR LA CRÉDIBILITÉ : vérifier de manière indépendante l'information auprès de sources crédibles telles que des experts ayant une expérience pratique dans le domaine. Ce qui est décrit a-t-il du sens d'après vos connaissances ?

Utiliser les techniques des **EXPERTS MÉDIA** pour compléter le travail d'investigation traditionnel et la recherche documentaire. Parmi ces techniques, on peut citer : la collecte de données, la géolocalisation, la reconnaissance biométrique, l'analyse de modèles et bien d'autres encore.

APPRENDRE ET S'ADAPTER : Les auteurs de fausses informations et de désinformations générées par l'IA s'adaptent en permanence. Ajustez votre approche à l'évolution du paysage.

PARTAGEZ VOTRE TRAVAIL

En plus d'une affirmation démentie, les experts recommandent de fournir une analyse transparente et des liens vers les sources. Cela peut aider les lecteurs à suivre une enquête et à comprendre un récit nuancé. Dans certains cas, l'enquête est plus importante que si le contenu est rédigé par une IA.

*** Henk van Ess est un expert en OSINT et en techniques de vérification des faits.*

EFCSN

Guide de référence rapide : Les choses à faire, à ne pas faire et les indices

Les indices suivants peuvent indiquer qu'un contenu a été généré par l'IA ou modifié numériquement. Associés aux autres conseils mentionnés dans ce guide (contexte, techniques d'investigation et outils de détection), ils peuvent vous aider à comprendre la vérité qui se cache derrière ce que vous voyez.

Texte

- A souvent (mais pas toujours) une **meilleure grammaire** qu'un humain.
- Susceptible d'utiliser un **langage trop formel ou structuré**, en particulier dans un contexte de réseaux sociaux.
- **Adverbes ou adjectifs excessifs.**
- Manque d'émotions humaines, d'humour, de sarcasme et d'expressions idiomatiques.
- Peut **manquer de détails spécifiques** (noms, dates, lieux) ou d'idées originales.
- Le plus important : les faits énoncés dans le texte sont-ils corrects ?

Vidéo

- N'utilisez pas un détecteur d'images généré par l'IA sur des images fixes tirées d'une vidéo.
- Examinez **les expressions faciales et les mouvements**, tels que les clignements d'yeux, et vérifiez si le mouvement de la bouche correspond à l'audio.
- Peuvent être marquées par **des transitions ou des coupures nettes.**

Audio

- **Comparez l'audio suspect à un échantillon authentique** à l'aide d'outils capables de détecter les différences dans les modèles de parole et de respiration, l'intonation...
- Lorsque vous utilisez des détecteurs, évitez les échantillons audio de faible qualité contenant des parasites ou des bruits de fond.
- Peut être marqué par des **schémas d'élocution non naturels ou mécaniques**, l'absence de pauses ou de respiration naturelle.

Images

- Recherchez les zones présentant des **détails non naturels** : peau parfaite, arrière-plan flou, beauté ou lumière non naturelle, et bizarreries telles que des doigts supplémentaires.
- **Recherchez un filigrane** dans les générateurs d'images courants.
- Prêtez attention aux détails de ce qui est représenté : est-ce logique ? Est-ce approprié ?
- Lorsque vous utilisez des détecteurs, **préférez une image en haute définition ou une version téléchargée en premier plutôt qu'une image qui a été partagée et re-partagée.**