Une étude de l'UC3M analyse les caractéristiques des deepfakes générés par l'IA

La plupart des deepfakes (vidéos avec de fausses recréations hyperréalistes) générés par l'intelligence artificielle (IA) qui se répandent sur les réseaux sociaux mettent en scène des hommes politiques et des artistes et sont souvent liés à des cycles d'actualité. Il s'agit de l'une des conclusions d'une étude de l'Universidad Carlos III de Madrid (UC3M) qui analyse les caractéristiques formelles et de contenu des désinformations virales en Espagne résultant de l'utilisation d'outils d'IA à des fins illicites. Ce progrès représente un pas en avant vers la compréhension et l'atténuation des menaces posées par les canulars dans notre société.

Dans l'étude, récemment publiée dans la revue OberCom, l'équipe de recherche a étudié ces faux contenus à travers les vérifications d'organisations espagnoles de fact-checking, telles que EFE Verifica, Maldita, Newtral et Verifica RTVE. « L'objectif était de déterminer une série de modèles et de caractéristiques communs dans ces deepfakes viraux, de fournir des clés pour leur identification et de faire des propositions pour l'éducation aux médias afin que les citoyens puissent faire face à la désinformation », explique l'une des auteures, Raquel Ruiz Incertis, chercheuse au département de communication de l'UC3M, où elle prépare un doctorat en communication européenne.

Les chercheurs ont développé une typologie de deepfakes, ce qui facilite leur identification et leur neutralisation. Selon les résultats de l'étude, certains dirigeants politiques (tels que Trump ou Macron) ont été les principaux protagonistes de contenus faisant référence à la consommation de drogues ou à des activités moralement répréhensibles. Il existe de plus une proportion considérable de deepfakes à caractère pornographique qui portent atteinte à l'intégrité des femmes, exposant notamment des chanteuses et des actrices célèbres. Ils sont généralement partagés à partir de comptes non officiels et se diffusent rapidement via les services de messagerie instantanée, soulignent les chercheurs.

La prolifération des deepfakes ou l'utilisation fréquente d'images, de vidéos ou de fichiers audio manipulés à l'aide d'outils d'IA, est un sujet brûlant d'actualité. « Ce type de canulars préfabriqués est particulièrement nuisible dans des situations délicates, par exemple à l'approche d'élections ou en période de conflit, tel qu'il est le cas actuellement en Ukraine ou à Gaza. C'est ce que nous appelons les « guerres hybrides » : la guerre n'est pas seulement menée dans le domaine physique, mais aussi dans le domaine numérique, et les mensonges sont plus répandus que jamais », explique Mme Ruiz Incertis.

Les applications de ces recherches sont diverses, allant de la sécurité nationale à l'intégrité des campagnes électorales. Les résultats suggèrent que l'utilisation proactive de l'IA sur les plateformes de réseaux sociaux pourrait révolutionner la façon dont nous maintenons l'authenticité des informations à l'ère numérique.

L'étude souligne le besoin d'une plus grande éducation aux médias et propose des stratégies éducatives pour améliorer la capacité du public à discerner le contenu réel du contenu manipulé. « Bon nombre de ces deepfakes peuvent être identifiés grâce à la recherche inversée d'images sur des moteurs de recherche tels que Google ou Bing. Des outils existent pour permettre aux citoyens de vérifier la véracité du contenu en deux clics avant de diffuser un contenu de provenance douteuse. La clé est de leur apprendre à le faire », indique Raquel Ruiz Incertis. Elle fournit de plus d'autres conseils pour détecter les deepfakes, tel que l'examen de la netteté des bords des éléments et la définition de l'arrière-plan de l'image : si les mouvements sont ralentis dans les vidéos ou s'il y a une quelconque altération du visage, une disproportion du corps ou un étrange jeu d'ombres et de lumières, tout porte à croire qu'il peut s'agir d'un contenu généré par l'IA.

De plus, les auteurs de l'étude estiment également qu'une législation est nécessaire pour obliger les plateformes, les applications et les programmes (tels que Midjourney ou Dall-e) à établir un « filigrane » les identifiant et permettant à l'utilisateur de savoir en un simple coup d'œil que cette image ou vidéo a été modifiée ou créée entièrement par l'IA.

L'équipe de recherche a utilisé une approche multidisciplinaire, combinant la science des données et l'analyse qualitative, pour examiner comment les organisations de vérification des faits appliquent l'IA dans leurs opérations. La méthodologie principale est une analyse de contenu d'une trentaine de publications extraites des sites Web des fact-checkers susmentionnés, où ces contenus manipulés ou fabriqués par l'IA sont réfutés.

Référence bibliographique: Garriga, M., Ruiz-Incertis, R., & Magallón-Rosa, R. (2024). Artificial intelligence, disinformation and media literacy proposals around deepfakes. Observatorio (OBS*), 18(5). https://doi.org/10.15847/obsOBS18520242445

Vidéo: https://youtu.be/pbU5UiYVsIE