



LES DÉTECTEURS D'IA fonctionnement et limites

🔴 Images surprenantes sur les réseaux sociaux, commentaires qui semblent écrits par des robots, textes ou voix artificiels de plus en plus convaincants... Face à la multiplication des contenus générés par intelligence artificielle (IA), de nombreux outils promettent de les identifier. En pratique, ils reposent aujourd'hui sur deux approches principales : **le marquage du contenu** et **l'analyse prédictive**. Dans ce nouvel épisode de la **Minute numérique**, faisons le point sur ces **outils de détection, leur fonctionnement, mais aussi leurs limites**.

🔑 **La première famille d'outils** consiste à ajouter une signature dans le fichier permettant d'indiquer qu'un contenu a été créé ou modifié par une IA. **C2PA (Content Credentials)** fonctionne comme une étiquette numérique. Des métadonnées signées cryptographiquement, c'est-à-dire des **informations invisibles intégrées au fichier**, permettent d'indiquer son origine et les principales modifications qu'il a subies. Lorsqu'elles sont présentes, elles constituent une indication fiable sur la provenance du contenu. Le principal problème est que ces métadonnées peuvent être supprimées très facilement. Une simple capture d'écran, une réexportation du fichier ou certaines plateformes de publication suffisent parfois à les faire disparaître. **L'outil de vérification** est disponible à cette adresse : <https://verify.contentauthenticity.org/>

🔍 **SynthID**, développé par Google DeepMind, utilise une approche différente : le marquage est directement intégré au contenu lui-même. **Il peut être dissimulé dans les pixels d'une**

image, dans un signal audio ou, pour certains textes, dans les choix de mots effectués par l'IA. Cette technique résiste généralement mieux aux transformations et aux copies que les simples métadonnées, mais elle n'est pas infaillible. **Les compressions, les modifications successives ainsi que les reformulations ou les traductions** peuvent fortement dégrader, voire faire disparaître, le marquage. Pour vérifier la présence de ce type de marquage, il faut généralement utiliser les outils proposés par le service qui a généré le contenu, lorsqu'ils existent.

 **La présence d'un marquage (C2PA ou SynthID) constitue généralement une indication fiable sur la provenance d'un contenu, mais son absence ne permet pas de conclure qu'il n'a pas été généré par IA.** Une personne souhaitant masquer l'origine d'un contenu peut facilement supprimer ou altérer ce marquage, **voire recourir à un générateur d'IA qui n'en utilise aucun.**

 **La seconde famille d'outils ne recherche aucun marquage. Elle analyse directement le contenu afin d'estimer la probabilité qu'il ait été produit par une IA.** Pour les textes, ces systèmes étudient notamment le vocabulaire, la structure des phrases ou certaines régularités stylistiques. Pour les images, ils recherchent des caractéristiques fréquemment associées aux générateurs d'images. Pour l'audio, ils analysent la voix ou le signal sonore afin de détecter d'éventuelles traces de synthèse. **Contrairement aux systèmes de marquage, ces outils ne fournissent qu'une estimation statistique, jamais une certitude.**

 **Ils produisent donc inévitablement des erreurs.** Pour les textes en particulier, **les faux positifs** peuvent toucher davantage les personnes qui écrivent dans une langue qui n'est pas leur langue maternelle ou dont le style d'écriture est atypique.

 **Il faut également garder à l'esprit qu'en pratique, la plupart des outils de vérification sont proposés sous la forme de services en ligne.** Pour analyser un texte, une image, un fichier audio ou une vidéo, il est généralement nécessaire d'envoyer ce contenu à un serveur exploité par l'éditeur du service. Cette étape peut soulever **des questions de confidentialité, de protection des données personnelles ou de secret professionnel**, notamment lorsque les contenus analysés ne sont pas destinés à être rendus publics.

 **L'esprit critique demeure donc indispensable :** avant de conclure qu'un contenu a été généré par une IA, ou au contraire qu'il est authentique, **il est préférable d'examiner sa source, son contexte et de croiser les informations disponibles.**

 Si vous rencontrez des difficultés dans la manipulation des outils numériques, **la CCAPV propose des ateliers numériques d'accompagnement** gratuits et accessibles à tous, y compris aux professionnels, organisés sur tout le territoire de la CCAPV.

 Contactez le **Conseiller numérique de la CCAPV** au 07 88 85 43 71 ou par courriel : conseiller.numerique@ccapv.fr

 Pour retrouver toutes les dates d'ateliers, rendez-vous sur le **site de la CCAPV** à cette adresse : ccapv.fr/services-publics/france-services/conseils-numeriques