Informations sur les **Deepfakes**

Autorité
luxembourgeoise
indépendante de
l'audiovisuel

Version du 12.01.2026

Autorité luxembourgeoise indépendante de l'audiovisuel (ALIA)

18, rue Erasme
L-1468 Luxembourg
+352 247-70 105
samra.cindrak@alia.etat.lu

Qu'est-ce qu'un deepfake ?

Les **deepfakes** (encore appelés **hypertrucages**) désignent des contenus d'un réalisme saisissant, créés ou modifiés par l'IA, capables d'imiter l'apparence, la voix ou les gestes d'une personne afin de lui faire dire ou faire des choses qu'elle n'a jamais dites ni faites. Au-delà de la simple imitation de personnes réelles, ces technologies permettent également de générer de toutes pièces des scènes ou des événements entièrement fictifs, de plus en plus réalistes. Les deepfakes sont des contenus synthétiques, conçus le plus souvent dans une intention de tromper ; ils visent à faire croire à des événements fictifs afin d'induire le public en erreur, d'influencer les perceptions et de causer un préjudice.

Grâce à l'intelligence artificielle (IA), il est aujourd'hui beaucoup plus simple, rapide et peu coûteux de manipuler ou de créer de grandes quantités d'images, de vidéos et d'audios. Des applications aux interfaces intuitives, souvent basées sur des logiciels open source, permettent aujourd'hui à un large public de générer facilement des vidéos truquées à partir d'un simple texte descriptif (« prompt »), sans nécessiter de compétences techniques approfondies. En démocratisant l'accès à ces technologies, ces évolutions ont accru le nombre de personnes en mesure d'en faire usage, augmentant ainsi le risque d'une utilisation malveillante et la portée potentielle de leurs effets nuisibles.

Malgré leur potentiel créatif, ces technologies peuvent être exploitées par des individus mal intentionnés pour :

Commettre des fraudes

- Par le biais de différentes méthodes de « **spoofing** », qui permettent à quelqu'un de se faire passer pour une autre personne ou une entreprise afin de gagner la confiance de ses victimes et de les tromper, manipuler ou escroquer. Les escroqueries dites « **CEO scams** » (ou « fraudes au président ») améliorées par deepfake, dans lesquelles des criminels imitent des dirigeants d'entreprise pour amener des employés à divulguer des données sensibles ou à effectuer des transferts d'argent, sont en augmentation nette. Ainsi, début 2024, un employé à Hong Kong a transféré 23 millions d'euros à des escrocs qui avaient simulé une visioconférence complète grâce à l'IA, lui faisant croire qu'il parlait à son responsable hiérarchique et au conseil d'administration de l'entreprise.
- Par le contournement de **systèmes de vérification biométrique** (reconnaissance faciale ou vocale, par exemple) à l'aide des deepfakes, permettant aux fraudeurs d'accéder à des smartphones, ordinateurs, applications, comptes bancaires et, par conséquent, à des données sensibles.
- **Par le recours aux deepfakes pour perfectionner des méthodes d'escroquerie classiques**, telles que « l'arnaque au petit-fils », les appels de choc, les arnaques sentimentales (« love scams ») ou d'autres formes de fraude en ligne. Les données des victimes potentielles peuvent être aisément compilées à partir de profils publics et analysées pour créer des profils précis ou des scripts de conversation personnalisés. Les messages peuvent être traduits sans erreur et diffusés à grand échelle. En outre, la possibilité de générer automatiquement des réponses personnalisées accroît

considérablement la scalabilité des activités frauduleuses, en permettant aux auteurs de mener de nombreux échanges en parallèle et d'atteindre ainsi un nombre bien plus important de personnes. La qualité croissante des deepfakes, combinée à une personnalisation accrue des messages, contribue à accroître la crédibilité de ces tentatives de fraude. Beaucoup de victimes se sentent rassurées lorsqu'elles croient parler en temps réel à une personne familière, dont elles reconnaissent la voix ou le visage lors d'un appel téléphonique ou vidéo, sans se douter qu'il s'agit d'une imitation. Une fois la confiance établie, elles partagent souvent davantage d'informations sensibles. Les escrocs peuvent alors tisser des liens émotionnels et exploiter ces relations pour perpétrer d'autres crimes tels que le « **grooming** » (approche en ligne d'un mineur à des fins d'exploitation sexuelle), le **cyberharcèlement** ou la « **sextorsion** » (chantage à partir d'images intimes).

- Dans ce contexte, les techniques « **d'ingénierie sociale** » exploitent les biais cognitifs et les émotions pour rendre crédibles des récits trompeurs. Appliquées aux deepfakes, elles renforcent et accélèrent toutes les formes d'arnaques en :
    - créant un sentiment d'urgence ou d'angoisse qui court-circuite la réflexion critique ;
    - usurpant l'apparence, la voix ou l'autorité de personnes reconnues pour susciter confiance et obéissance ;
    - jouant sur la bienveillance ou l'empathie des victimes pour les inciter à divulguer des informations ou à transférer des fonds.

Discréditer des particuliers ou des adversaires politiques

Les deepfakes peuvent être utilisés de manière double : d'une part, pour discréditer ou diffamer une personne en lui prêtant de faux propos ou comportements, et d'autre part, pour exploiter la crédibilité et l'image de figures publiques reconnues afin de donner plus de poids et de légitimité à des messages trompeurs. Les personnalités politiques figurent souvent parmi les premières cibles de campagnes de manipulation, mais d'autres figures publiques, telles que des acteurs célèbres, des animatrices connues ou encore des marques médiatiques populaires, sont également concernés en raison de leur forte visibilité et de la confiance qu'elles inspirent. Comme il existe beaucoup de matériel audiovisuel public à leur sujet, il est facile d'alimenter les IA avec ces données pour produire des deepfakes très réalistes.

Désinformer et manipuler

Les deepfakes servent aussi à produire et propager des campagnes coordonnées de désinformation visant à :

- manipuler l'**opinion publique** ;
- semer le doute et miner la confiance envers les **institutions publiques**, les **médias** et la **démocratie** en général ;
- porter atteinte à l'intégrité des **élections** et des processus démocratiques ;
- diviser la **société**.

Les contenus générés par l'IA peuvent être utilisés de multiples manières à des fins de manipulation, en particulier à l'approche des élections, et représentent ainsi un risque sérieux pour les processus démocratiques.

Ainsi, deux jours avant les élections législatives en Slovaquie, en septembre 2023, a circulé un enregistrement audio généré par IA dans lequel le dirigeant libéral de l'opposition, Michal Šimečka, semblait admettre avoir tenté d'influencer le scrutin en faveur de son parti. Peu avant la primaire du New Hampshire, aux États-Unis, en janvier 2024, un appel automatisé (robocall) imitant la voix de Joe Biden a été diffusé, dans le but apparent de dissuader des électeurs de se rendre aux urnes. Déjà en avril 2023, le Parti républicain avait publié, immédiatement après l'annonce de la candidature de Biden à un second mandat, une vidéo de campagne générée par IA décrivant un scénario catastrophe fictif de sa réélection : scènes de chaos, effondrement économique, guerre et migration incontrôlée, destinées à dissuader l'électorat. L'exemple de Rumeen Farhana, femme politique de l'opposition au Bangladesh, dont des images générées par IA la montrant en bikini ont circulé fin 2023, illustre qu'un préjudice considérable peut survenir même lorsque le trucage est manifeste. Bien que la manipulation ait été rapidement reconnue, ces images ont suscité une vive indignation dans un pays majoritairement musulman. Et puis il y a Donald Trump, qui diffuse régulièrement des deepfakes le mettant en scène, dont récemment un deepfake le représentant larguant des excréments depuis un avion de chasse sur les manifestants du mouvement "No Kings", afin de tourner en dérision les opposants à son gouvernement qui protestaient contre sa dérive autoritaire et son emprise grandissante sur le pouvoir.

Des acteurs malveillants recourent aux deepfakes pour **influencer les choix électoraux de diverses manières** : en portant atteinte à l'intégrité des adversaires politiques, en dissuadant certains électeurs de voter ou en diffusant délibérément de fausses informations électorales. Ces technologies sont aussi exploitées pour mettre en scène des scénarios de peur ou de menace, en transformant des idées abstraites en images concrètes, rendant ainsi les messages véhiculés plus crédibles, plus saisissants et plus persuasifs. Plus largement, elles servent à orienter l'opinion publique, notamment en mobilisant contre des thèmes sensibles tels que la migration ou l'accueil des réfugiés, dans le but de renforcer certains agendas politiques. Dans le même temps, elles contribuent à consolider et amplifier des narratifs pro-gouvernementaux, tout en détournant l'attention de sujets critiques à l'égard du pouvoir et en saturant l'espace informationnel d'informations non pertinentes, trompeuses ou fausses.

La prolifération de la désinformation et la possibilité de créer, grâce à l'IA, des contenus indiscernables du réel font que les **individus perdent une faculté humaine fondamentale** : celle de **comprendre** le monde à travers leurs **sens**, de se fier à ce qu'ils voient ou entendent pour en tirer du sens. Ce brouillage du rapport au réel engendre une perte de confiance, le sentiment que plus rien n'est vrai ni vérifiable, et fragilise ainsi la confiance collective. Une telle situation risque d'être amplifiée par ce que la recherche désigne sous le terme de « **dividende du menteur** » (**liar's dividend**), qui permet à des acteurs politiques de remettre en cause, voire nier, l'authenticité de faits avérés pouvant leur être défavorables, en invoquant la possibilité de manipulations par deepfake. L'usage des deepfakes devient alors un instrument de déni stratégique, contribuant à l'érosion de la confiance dans les médias, les institutions et la notion même de vérité.

Au **Luxembourg**, les partis DP, LSAP, déi gréng, ADR, Déi Lénk, Piraten, Fokus et Volt ont signé en janvier 2023 un **accord électoral** par lequel ils s'engagent à mener des campagnes équitables et factuelles dans le cadre des élections européennes de 2024. Ils y affirment leur

volonté d'une utilisation responsable des réseaux sociaux et renoncent aux attaques personnelles, à la diffusion de fausses informations, aux campagnes de diffamation ainsi qu'à la détérioration ou manipulation du matériel électoral d'autres partis. Si l'accord ne fait pas explicitement référence à une renonciation à l'usage des deepfakes, il évoque néanmoins l'engagement à ne pas recourir à des bots ni à des campagnes automatisées de propagande ou de manipulation. On peut dès lors raisonnablement en déduire que de telles technologies IA ne devraient pas être utilisées à des fins de désinformation.

Produire des contenus illégaux

Cela inclut notamment les « **deepnudes** », c'est-à-dire des contenus **pornographiques** générés sans le consentement des personnes concernées.
La production, la possession ou la diffusion d'images pédopornographiques, qu'elles soient générées par IA ou non, mettant en scène des mineurs est strictement interdite et punissable par la loi.

Combiner plusieurs tactiques malveillantes

Les deepfakes sont rarement diffusés de manière isolée : ils s'intègrent souvent dans des **campagnes hybrides** combinant diverses tactiques basées sur l'IA pour en maximiser la portée et l'efficacité.

Au **Luxembourg**, des deepfakes ont circulé, mettant en scène notamment le Premier ministre Luc Frieden, la bourgmestre de la Ville de Luxembourg, Lydie Polfer, le Grand-Duc sortant, Henri, ainsi que l'ancien Premier ministre Jean-Claude Juncker. Dans ces vidéos et articles manipulés ou générés par l'IA, les personnalités donnaient des conseils financiers et incitaient le public à investir auprès de plateformes de cryptomonnaies. Des journalistes de RTL (Mariette Zenners, Caroline Mart et Lynn Cruchten), ainsi que l'actrice Désirée Nosbusch, ont également été victimes de deepfakes afin de rendre ces arnaques plus crédibles.

Les escrocs ont combiné plusieurs méthodes :

- reproduction de sites d'actualités sérieux et connus (RTL Lëtzebuerg, « tagesschau ») avec logos, couleurs et mises en page copiés ;
- création de faux articles à titres accrocheurs promettant des révélations sensationnelles ;
- promotion de ces articles au moyen de faux profils et en recourant aux services publicitaires payants de Facebook et Instagram.

Que prévoit la loi ?

La loi impose des mesures à plusieurs niveaux :

- aux **développeurs** de modèles d'IA générative ;
- aux **plateformes** où ces contenus sont diffusés ;
- aux **utilisateurs** (**déployeurs**) de ces outils lorsqu'ils n'utilisent pas l'IA dans un cadre strictement privé.

Les développeurs d'IA doivent :

- garantir que toute personne interagissant avec un système d'IA puisse reconnaître qu'il s'agit d'une machine et non d'un humain ;
- s'assurer que tout résultat produit ou modifié par l'IA soit identifiable comme tel par un marquage lisible par machine. L'ajout de filigranes invisibles ou de métadonnées (date de création ou de modification, modèle utilisé, etc.) constituent des exemples de marquages lisibles par machine.

Des mesures préventives sont nécessaires pour prévenir la production de deepfakes nuisibles, telles que :

- la mise en place de filtres d'entrée et de sortie pour prévenir la création de deepfakes nuisibles (notamment ceux visant des mineurs, les deepnudes, etc.) ;
- des tests réguliers de vulnérabilité des systèmes d'IA générative, ainsi que des tests de vérification de la légalité des contenus produits.

Les plateformes doivent :

- établir des règles claires sur la création et le partage de contenus synthétiques, dans le cadre de leur obligation de modération des contenus illicites ;
- prévoir des sanctions (suppression de comptes, etc.) pour les utilisateurs contrevenants.
- permettre le signalement des deepfakes qui sont susceptibles de constituer un contenu illicite.

Les plateformes sont encouragées à :

- fournir des outils permettant d'indiquer si un contenu a été modifié ou généré par IA ; notamment par la mise en place d'une « notice sur l'IA » sous les publications concernées.

Les utilisateurs doivent :

- indiquer clairement lorsqu'ils génèrent un deepfake et s'abstenir de générer des deepfakes nuisibles ou illégaux ;
- signaler tout contenu suspect ou trompeur.

Comment reconnaître un deepfake ?

Il existe de moins en moins d'indices visuels ou sonores fiables pour identifier un deepfake : les modèles deviennent si performants que les erreurs évidentes disparaissent. Souvent, seules des analyses via un logiciel spécialisé ou des examens manuels poussés et détaillés (image par image, ralenti, etc.) permettent de les détecter, sans pour autant garantir, à ce stade, une fiabilité absolue.

De nombreuses personnes croient à tort qu'un contenu est authentique simplement parce qu'elles ne détectent pas de « faute typique » d'IA.

La détection repose donc souvent sur le **contexte**, le **contenu** et les **incohérences logiques** qui y sont relatifs. L'IA sait imiter des voix et des gestes, mais comprend difficilement les lois de la physique ou la causalité : un comportement illogique dans une vidéo (p. ex. quelqu'un reste immobile alors qu'il y a un incendie ; une voiture à contre-sens sur une piste cyclable) peut rompre l'illusion d'un deepfake.

Que pouvez-vous faire ?

- **Restez informé !** Lisez des articles, abonnez-vous à des newsletters ou podcasts sur l'IA. La technologie évolue vite : mieux vous la comprenez, plus vous repérez facilement les faux.
- **Vérifiez la présence d'un label IA.** La manière dont le label est conçu et positionné (couleur, texte, symbole, emplacement) peut varier en fonction des plateformes et de leurs systèmes d'affichage. Souvent, la mention précisant qu'il s'agit d'un deepfake figure dans la légende de l'image ou de la vidéo. Cherchez des mentions comme #AI, #GeneratedWithAI ou #deepfake, ou un logo de l'outil utilisé (par exemple, Sora d'OpenAI).
- **Ne partagez pas à la légère.** Chaque like, commentaire ou partage contribue à la diffusion de ces faux contenus. Si vous doutez de l'authenticité, abstenez-vous de partager.
- **Vérifiez la couverture médiatique.** Les médias sérieux traitent rapidement les événements importants. S'il n'y a aucune trace d'une nouvelle ailleurs, il s'agit probablement d'un faux. Il est toujours préférable de vérifier l'information en consultant plusieurs sources plutôt que de se fier à une seule. Pensez aussi à visiter des sites de vérification comme Fact-checks – EDMO Belux ou RTL Infos – Fact check.
- **Adoptez un esprit critique face à certains contenus.** On estime qu'en 2026, près de 90 % du contenu en ligne sera généré par l'IA. Restez prudent, mais sans tomber dans une méfiance généralisée envers les sources officielles et journalistiques fiables. Dans un monde où la frontière entre réalité et fiction se brouille de plus en plus, la confiance dans une communication transparente, responsable et fiable est essentielle.

Liens et ressources utiles

- Retrouvez des informations essentielles et des conseils pratiques pour vous protéger contre la fraude financière sur letzfin.lu, la plateforme de référence en matière d'éducation financière de la Commission de Surveillance du Secteur Financier (CSSF). La rubrique Précautions à prendre sensibilise le public aux arnaques financières les plus fréquentes, notamment le phishing assisté par IA, l'usurpation d'identité et les faux sites web. Suivez aussi @letzfin sur Instagram pour rester informés.
- Le Centre National de Compétences en Cybersécurité (NC3) hébergé au Luxembourg House of Cybersecurity (LHC), propose aux entreprises et institutions des conseils et des outils pratiques pour renforcer leur sécurité numérique, y compris face aux attaques exploitant l'intelligence artificielle.
- En cas de fraude en ligne, le site cyberfraud.lu vous oriente vers les bonnes démarches et les interlocuteurs compétents. Lancée en juin 2025 par le Luxembourg House of Cybersecurity (LHC) et l'Association des banques et banquiers Luxembourg (ABBL), dans le cadre d'une campagne nationale de sensibilisation, cette initiative est soutenue par plus de quinze partenaires.
- Pour toute question générale sur la cybersécurité ou la sécurité en ligne, contactez la **BEE SECURE Helpline** au **+352 8002 1234** ou via le formulaire en ligne.