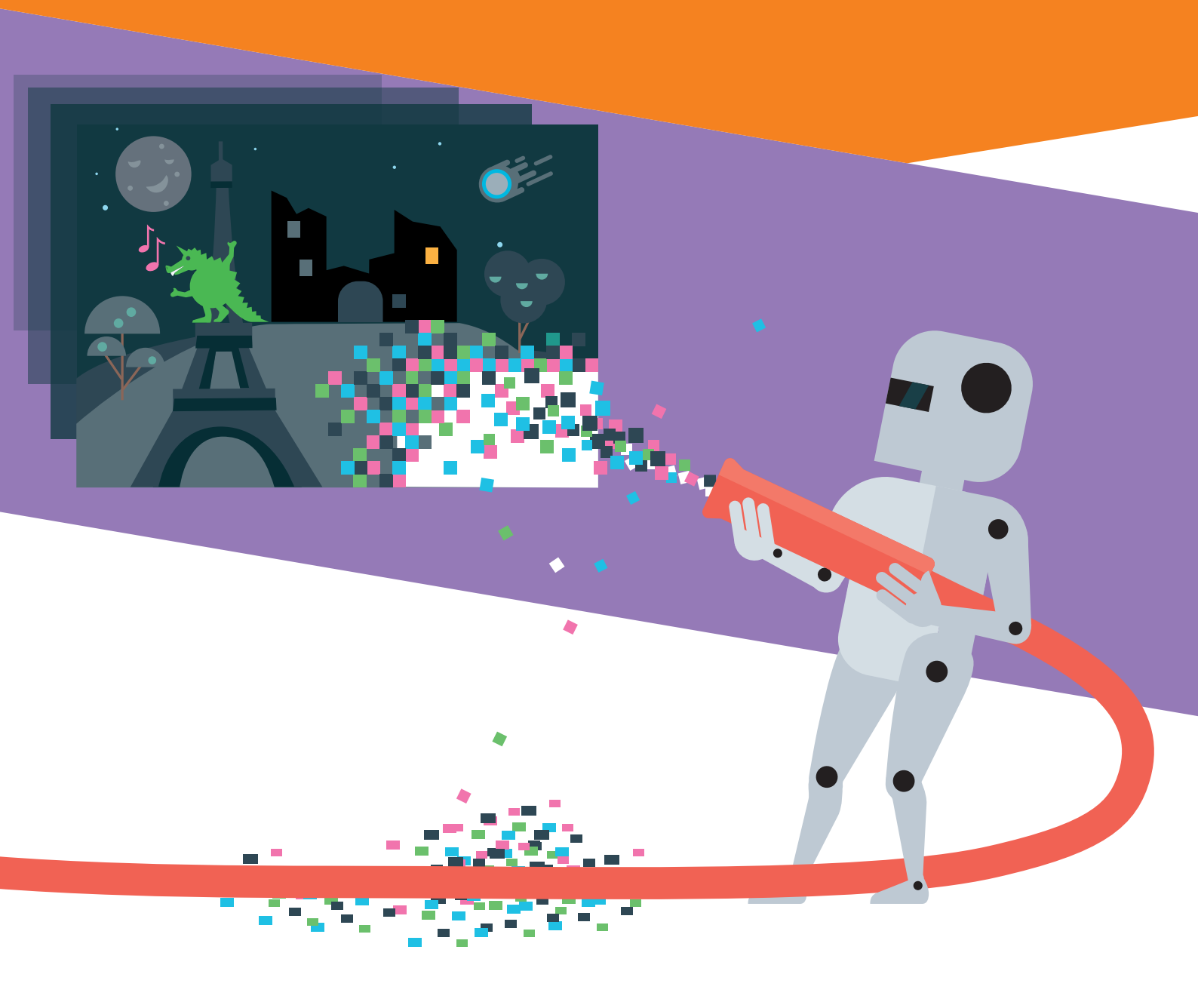
Nos yeux et nos oreilles à l'épreuve

Synthèse de l'étude de TA-SWISS « Deepfakes et réalités manipulées »



TA-SWISS, Fondation pour l'évaluation des choix technologiques et centre de compétence des Académies suisses des sciences, entend mener une réflexion sur les répercussions – opportunités et risques – de l'utilisation de nouvelles technologies.

La synthèse se base sur une étude scientifique réalisée pour le compte de TA-SWISS par un groupe de projet interdisciplinaire, sous la direction générale de urat Karaboga (Fraunhofer-Institut für System- und Innovationsforschung ISI à Karlsruhe). Les personnes suivantes y ont contribué : Nula Frei (Institut de droit européen, Université de Fribourg), Manuel Puppis et Patric Raemy (Département des sciences de la communication et de la recherche sur les médias DCM, Université de Fribourg), Daniel Vogler (Forschungszentrum für Öffentlichkeit und Gesellschaft fög, Université de Zurich), Frank Ebbers (Competence Center Neue Technologie au Fraunhofer ISI, Karlsruhe), Greta Runge (Fraunhofer-Institut für System- und Innovationsforschung ISI, Karlsruhe), Adrian Rauchfleisch (Graduate Institute of Journalism, National Taiwan University), Gabriele de Seta (Department of Linguistic, Literary and Aesthetic Studies, Université de Bergen), Gwendolyn Gurr (analyste de données d'audience, Schweizer Radio und Fernsehen SRF), Michael Friedewald (Fraunhofer-Institut für System- und Innovationsforschung ISI, Karlsruhe), Sophia Rovelli (Institut de droit européen, Université de Fribourg).

Cette synthèse présente les principaux résultats et les recommandations de l'étude sous forme condensée et s'adresse à un large public.

Synthèse de l'étude de TA-SWISS « Deepfakes et réalités manipulées »

Murat Karaboga, Nula Frei, Manuel Puppis, Daniel Vogler, Patric Raemy, Frank Ebbers, Greta Runge, Adrian Rauchfleisch, Gabriele de Seta, Gwendolyn Gurr, Michael Friedewald, Sophia Rovelli

TA-SWISS, fondation pour l'évaluation des choix technologiques (éd.)
vdf Hochschulverlag an der ETH Zürich, 2024.

ISBN : 978-3-7281-4185-9

L'étude est également disponible en open access : www.vdf.ch

La synthèse peut être téléchargée gratuitement : www.ta-swiss.ch

Les deepfakes en bref

L'intelligence artificielle (IA) continue de se développer à un rythme effréné, tout comme la production de films, d'images et d'enregistrements audio « de synthèse » qui ne témoignent pas de faits réels mais sont générés par des programmes informatiques. Comme ces deepfakes sont toujours plus faciles à produire, leur importance dans notre société est destinée à croitre rapidement. Or, s'ils offrent des opportunités dans le secteur du divertissement, de l'éducation et de la formation, les risques qu'ils présentent ne sont pas à négliger – notamment en termes politiques, de harcèlement moral ou de délits économiques.

Les deepfakes – ou médias synthétiques – sont des images, des vidéos ou des enregistrements sonores générés par une intelligence artificielle et montrant une situation qui n'a jamais existé sous cette forme. Il peut s'agir de contenus manipulés, ou même entièrement artificiels, créés par des logiciels qui s'appuient sur des données d'entraînement provenant d'immenses bases de données sur internet. La gamme des programmes de deepfake qui se sont imposés aujourd'hui est vaste, comprenant des logiciels simples à utiliser pour l'échange de visages *(face swapping)* comme des applications sophistiquées permettant un « jeu de marionnettes virtuel » avec des personnes artificielles *(full body puppetry).* Il existe aussi déjà des logiciels, pour l'instant encore rudimentaires, qui peuvent créer des vidéos à partir de commandes textuelles *(prompts).*

Quelques opportunités …

Les médias synthétiques ont un véritable potentiel pour l'industrie du divertissement. Leurs applications à des fins économiques sont également prometteuses, par exemple pour la présentation de vêtements ou autres produits par des influenceurs artificiels. Dans les écoles, les cours d'histoire pourraient aussi gagner en attractivité grâce à des avatars de personnalités du passé – Jules César, Catherine la Grande, Napoléon – qui s'entretiendraient de manière interactive avec les élèves. Quant aux autorités, elles espèrent en tirer profit dans le cadre d'enquêtes pour visualiser le déroulement des faits lors d'investigations criminelles.

… et quelques risques

Un deepfake malveillant peut montrer des personnes commettant des actes répréhensibles qu'elles n'ont jamais commis ou prononçant des mots qu'elles n'ont jamais dits. De telles vidéos ou enregistrements audio truqués peuvent servir à faire chanter ou à compromettre des personnes – un procédé utilisé dans le cadre de tensions ou de conflits politiques – ou à faire des ravages dans les relations privées, par exemple sous la forme de faux contenus pornographiques *(revenge porn).*

Il est aussi possible de cloner la voix d'une personne à des fins frauduleuses, pour obtenir de l'argent de son cercle d'amis ou de sa famille, ou la voix d'un supérieur hiérarchique pour commettre d'autres délits économiques comme l'espionnage de secrets commerciaux.

Ces procédés posent un défi majeur aux médias qui doivent investir des ressources considérables pour s'assurer de l'authenticité des vidéos et éviter de contribuer eux-mêmes à la propagation de deepfakes.

Recommandations principales

Face à l'évolution rapide de la technologie, seule une approche combinée de diverses mesures de protection permettra d'exploiter le potentiel des médias synthétiques tout en limitant leurs effets néfastes. Les mesures politiques, les détecteurs de deepfakes, le marquage des contenus de synthèse tel que l'envisagent les grands fournisseurs de logiciels et la sensibilisation médiatique aux deepfakes doivent se compléter mutuellement. La responsabilité personnelle est également un facteur important : il s'agit de considérer les vidéos sur internet avec un scepticisme sain et de télécharger avec retenue les photos et vidéos privées.

L'État devrait imposer aux grandes plateformes en ligne la suppression de tout deepfake préjudiciable aux individus. Étant donné que les particuliers sont généralement en position de faiblesse dans les conflits avec les grandes plateformes en ligne, il faut des services spécialisés pour conseiller et soutenir les victimes de deepfakes – ou les personnes concernées par des suppressions injustifiées. La Confédération et les cantons devraient doter les services d'aide aux victimes de cyberdélits de moyens suffisants.

Une vision déformée de la réalité

Nous considérons en général comme vrai ce que nous voyons de nos yeux et entendons de nos oreilles. La fonction représentative de la réalité des vidéos, en particulier, est rarement mise en doute – du moins jusqu'à récemment. En effet, la technologie permet aujourd'hui de produire à peu de frais des vidéos et enregistrements audio trompeusement réalistes d'événements qui n'ont jamais eu lieu.

Vers la fin de l'automne 2023, on a cru assister à l'avènement d'un futur top model. Emily Pellegrini postait sur son nouveau canal Instagram des vidéos qui rendaient ses admirateurs fous d'enthousiasme. Le nombre de ses followers se multipliait à toute allure. À la longue série d'emojis en forme de cœur et de flamme dans la colonne des commentaires s'ajoutaient les demandes de contact. Les journaux annoncèrent qu'un footballeur allemand insistait pour obtenir un rendez-vous et qu'un milliardaire, une star du tennis et d'autres grands noms du monde du sport tentaient également de séduire la belle – jusqu'à ce qu'ils se rendent compte qu'ils n'avaient pas jeté leur dévolu sur une femme en chair et en os, mais sur un avatar créé par une intelligence artificielle (IA). Selon le journal britannique Daily Mail, son créateur – resté anonyme – annoncera par la suite avoir pris pour modèle la « femme de rêve de l'homme moyen » pour créer son Emily artificielle qui lui aura rapporté 10 000 dollars par mois.

Emily Pellegrini, une *fun-loving girl* selon ses propres termes, qui présente ses atouts physiques sur des plateformes payantes comme Onlyfans et Fanvue, est une influenceuse d'un nouveau type : des personnages de synthèse, mais trompeurs par leur réalisme, principalement de sexe féminin, capables notamment de chatter avec leur public grâce à des générateurs de texte basés sur l'IA. Les atouts que comportent ces individus artificiels pour les publicitaires sont évidents : une fois créés, ils ne demandent pas de salaire horaire, ne se fatiguent jamais et obéissent à toutes les instructions.

Profonde illusion

Des logiciels basés sur l'intelligence artificielle, notamment sur des réseaux neuronaux artificiels, permettent de créer des vidéos montrant des faits qui ne se sont jamais produits dans la réalité. Il peut s'agir de catastrophes naturelles ou d'explosions qui n'ont jamais eu lieu. Ou de vidéos de personnalités connues qui disent ou font quelque chose qu'elles n'ont jamais dit ou fait. Ainsi, un petit film du vidéaste Bob de Jong montre l'ancien Premier ministre néerlandais Mark Rutte, un violon sous son double menton, jouant avec émotion la partie solo de « Douce nuit ». L'artiste d'Amsterdam met en ligne sa création sur Diep Nep, sa chaîne YouTube.

Traduit en anglais, diep nep devient deepfake, un terme qui s'est entretemps également imposé dans les pays francophones pour désigner des vidéos authentiques en apparence, mais en réalité fortement manipulées, voire entièrement créées par ordinateur. Les spécialistes parlent aussi de « médias synthétiques » lorsque les contenus sont fabriqués à partir de données disponibles sur internet, notamment d'images, de vidéos et d'enregistrements sonores provenant des médias sociaux et de plateformes vidéo. Mais tandis que Bob de Jong déclare ouvertement que ses créations sont des artefacts, la paternité de nombreux deepfakes – voire leur grande majorité – reste obscure. Dans le présent résumé, les termes « deepfake » et « médias synthétiques » sont considérés comme synonymes.

Des œuvres pionnières de l'ombre à la lumière

Les premiers trucages vidéo apparaissent à l'automne 2017 sur Reddit, une sorte de réservoir virtuel de contenus issus des médias sociaux. Des petits films sont mis en ligne sous l'appellation « DeepFake », où le visage de l'actrice originale d'une vidéo pornographique est remplacé par les traits d'Emma Watson, de Gal Gadot ou d'autres stars de cinéma. Peu après, un autre utilisateur de Reddit met à disposition un logiciel appelé FakeApp, qui permet à n'importe qui de créer des deepfakes. Auparavant réservée aux studios hollywoodiens disposant de moyens financiers considérables, la production d'images de synthèse 3D nécessitant une forte puissance de calcul devient alors accessible à toute personne capable d'exécuter les cinq étapes de fabrication simples prescrites par FakeApp.

Par la suite, des images pornographiques manipulées sont téléchargées en masse sur le darkweb. Ces vidéos, au départ vite démasquées comme des trucages en raison de leur faible résolution et de leurs mouvements saccadés, deviennent toujours plus réalistes grâce aux progrès techniques. Il y a longtemps que les célébrités ne sont plus les seules à voir leur image détournée de la sorte. Au contraire, toute personne qui a un jour contrarié quelqu'un peut aujourd'hui en devenir la victime : depuis février 2018, l'expression revenge porn a sa propre entrée sur Wikipédia, qui mentionne d'ailleurs aussi la réalisation de faux films de nus.

Aujourd'hui encore, on estime que la grande partie des deepfakes téléchargés sont des représentations de pornographie féminine, même si le genre s'est entretemps étendu à d'autres domaines, notamment à la politique. En effet, les personnalités dont de nombreux enregistrements circulent sur internet offrent un matériau particulièrement riche pour réaliser des trucages vidéo.

En compétition pour le meilleur fake

Il existe une technique dans le domaine de la production de deepfakes qui attire beaucoup l'attention : on appelle réseaux adverses génératifs *(Generative Adversarial Networks,* ou GAN) les programmes informatiques qui peuvent créer des images similaires à leurs données d'entrainement, mais inédites. Ces logiciels comportent deux parties qui s'affrontent : le discriminateur et le générateur. Alors que ce dernier crée des images semblables à ses images d'entrainement, le discriminateur identifie les différences entre ces nouvelles images et les images originelles. Dans leur course l'un contre l'autre, générateur et discriminateur s'améliorent mutuellement, si bien qu'au final, ils produisent des images dont le style ne peut plus guère être différencié des modèles réels.

Une autre approche consiste à utiliser des encodeurs automatiques, c'est-à-dire des réseaux neuronaux artificiels capables de filtrer les caractéristiques essentielles d'un ensemble de données d'images et de les appliquer à d'autres images.

Des marionnettes avec un corps étranger

Le degré de sophistication de la manipulation d'un deepfake est plus ou moins élevé, mais cela ne reflète pas forcément l'impact qu'il aura sur le public.

Ainsi, les manipulations peuvent se limiter à modifier l'expression du visage et les mouvements de la bouche, c'est-à-dire à transférer les mimiques d'un acteur ou d'une actrice sur la personne cible. Dans le jargon, cette nouvelle mise en scène d'un visage est appelée *facial reenactment*. Lorsqu'une personne se synchronise elle-même sur une vidéo, c'est-à-dire lorsqu'elle fait correspondre les mouvements de sa bouche avec ses déclarations dans une autre langue,

la même approche s'applique. L'entreprise HeyGen Labs a par exemple développé un programme vidéo basé sur l'IA qui enregistre le son d'une vidéo, en traduit le contenu et retranscrit ce qui est dit dans l'autre langue dans la vidéo – et ce de sorte que la voix de la personne qui parle est clonée et que ses lèvres bougent en fonction de ce qui est dit. Cela permet aux journalistes et présentatrices de synchroniser leurs propres interventions.

Un autre type de deepfake, appelé *face morphing,* consiste à fusionner les traits de deux individus. Cette technique est principalement utilisée dans les milieux criminels pour falsifier des documents d'identité pour qu'ils puissent être utilisés par plusieurs personnes à la fois.

L'échange de visages, ou *face swapping,* a été utilisé dans la phase initiale des deepfakes décrite plus haut lors de la production de fausses vidéos pornographiques. Cette technique est également populaire à des fins de divertissement : des applications gratuites circulent sur internet qui permettent de « recréer » des scènes de films emblématiques en remplaçant par exemple le visage de Leonardo di Caprio ou celui de Kate Winslet par son propre portrait.

Les logiciels basés sur l'IA sont en outre capables de concevoir de toute pièce de nouvelles images de personnes qui n'existent pas. De tels portraits et vidéos produits par des générateurs de visages sont utilisés comme avatars dans les jeux vidéo ou comme interlocuteurs virtuels dans les services à la clientèle entièrement automatisés.

Enfin, il existe un logiciel qui agit comme un marionnettiste pour modifier les poses et les mouvements d'un individu dans une vidéo. Ce type de deepfake, appelé full body puppetry, est considéré comme le plus complexe. De même, il n'existe pas encore de kit d'IA complet capable de créer intégralement une vidéo avec animation vocale et synthèse de la voix. Il faut encore combiner entre eux plusieurs programmes, parfois payants et compliqués à utiliser, ce qui rend la création de vidéos deepfake d'apparence réelle difficile.

Il semble toutefois que la production de vidéos de synthèse sur simple pression d'un bouton est à portée de main : inspirés par les générateurs d'images assistés par l'IA capables de produire une image photoréaliste de licorne ou un faux tableau de Rembrandt à partir de commandes textuelles, les premiers programmes assistés par l'IA qui permettent de générer des vidéos deepfake sur la base de commandes vocales ou textuelles commencent à apparaitre. Il est vraisemblable que de tels logiciels seront accessibles au plus grand nombre dans un avenir proche.

Des voix sorties d'un laboratoire

Il n'y a pas que l'œil qui peut être trompé, l'oreille aussi : un logiciel est désormais capable de cloner la voix et les habitudes d'élocution d'une personne. En 2018, une entreprise écossaise réalise une expérience devenue célèbre dans les milieux spécialisés. Elle réussit à réanimer, du moins acoustiquement, le président américain assassiné John F. Kennedy : sur la base de son manuscrit et grâce à de nombreux enregistrements vocaux de l'homme d'État, elle produit une retranscription audio d'un discours qu'il n'a jamais pu prononcer à Dallas à l'automne 1963 en raison de l'attentat dont il a été victime. Son accent de Boston et sa cadence de parole sont parfaitement imités.

Ces dernières années, cette technologie a encore progressé. Pour modéliser la façon de parler d'une personne, il suffit aujourd'hui d'un ordinateur portable standard et de quelques secondes d'un clip audio – par exemple tiré d'une conférence téléchargée sur YouTube. Ce qu'on appelle la synthèse vocale a aussi beaucoup évolué et se retrouve dans des logiciels qui convertissent les écrits en clips audio. Ce procédé permet la production automatisée de livres audio et offre une solution aux personnes malvoyantes qui souhaitent écouter des textes lus à voix haute.

Outils d'identification des deepfakes

Le ton alarmant du terme « deepfake » révèle l'intention de nombreuses vidéos et enregistrements sonores de synthèse : induire le public en erreur et l'influencer en faveur de l'expéditeur. Le débat sur la manière de gérer ces leurres visuels et sonores est vif parmi les spécialistes.

Une des approches consiste à rendre transparente l'origine de tout enregistrement, ce qui pourrait se faire grâce à une signature numérique apposée directement sur un fichier vidéo ou audio pendant l'enregistrement. Si, basée sur la blockchain, cette solution semble relativement facile à mettre en œuvre sur un plan technique, une telle « empreinte digitale numérique » reste toutefois falsifiable, bien qu'au prix d'efforts considérables. Des voix critiques mettent aussi en garde contre le fait que cela donnerait aux régimes autoritaires et aux services secrets un instrument pour traquer les lanceurs d'alerte, les défenseurs des droits humains et les journalistes gênants. Par ailleurs, si une signature est capable d'attester de l'origine d'un enregistrement, cette certification peut aussi bien se rapporter à l'événement en entier qu'à une partie de l'action seulement. Enfin, cela ne permet pas non plus d'identifier les vidéos montrant des événements recréés avec des actrices et des acteurs qui, bien que techniquement « réelles », n'en diffusent pas moins potentiellement des contre-vérités. C'est le scénario repris par une série télévisée norvégienne dans laquelle un groupe de partisans montre une vidéo de l'assassinat – supposé – d'un premier ministre controversé, pour lui permettre de mieux se cacher.

Détecter les caractéristiques d'un trucage

D'autres approches visent à démasquer les vidéos ou les enregistrements audio artificiels en fonction de certaines caractéristiques. Mais en ce qui concerne le clonage vocal, tout le monde est d'accord : ces deepfakes acoustiques sonnent désormais tellement vrai qu'il est difficile de les distinguer de la voix d'une vraie personne, surtout lorsqu'ils sont diffusés au téléphone. Les autorités policières mettent donc en garde contre les faux appels choc qui fonctionnent d'autant mieux que la victime croit entendre la voix d'un membre de la famille lui demandant de l'argent.

Les trucages vidéo, quant à eux, peuvent être trahis par certains artefacts qui, s'ils ne sont pas nécessairement détectés par des yeux humains, le sont par des algorithmes basés sur l'IA. Les bords douteux, fondus non naturels, déformations et flous sont des indices de deepfake. Désormais, il existe des programmes de détection censés démasquer ces

fausses vidéos. Deux détecteurs gratuits ont été testés dans le cadre de l'étude de TA-SWISS. Ces tests ne se sont pas révélés satisfaisants, car les deux programmes ont donné de faux résultats. Un autre problème, et non des moindres, est que les détecteurs ont déclaré que certaines vidéos authentiques étaient des trucages – ce qui risque d'affecter la crédibilité de contenus originaux. Il faut s'attendre à ce que les développeurs de logiciels deepfake restent informés et mettent tout en œuvre pour ajuster leurs programmes et déjouer ces éléments révélateurs – un jeu du chat et de la souris dans lequel les faussaires auront sans doute l'avantage.

Maintenir une saine méfiance

Pour identifier un deepfake, le bon sens devrait faire jeu égal avec les détecteurs techniques. Un examen critique de la source et du contenu d'une vidéo, la vigilance face à des incohérences de détails tels que des mèches de cheveux, des doigts ou des boucles d'oreilles, ainsi que l'attention portée à un comportement inhabituel de la personne filmée peuvent aider à les démasquer. Il est aussi possible de s'entrainer à reconnaitre les vidéos truquées sur des sites web comme Detectfakes.

Un peu plus de suspicion aurait certainement évité au footballeur allemand et autres admirateurs d'Emily Pellegrini de tomber dans le panneau. En comparant différentes vidéos, on constate en effet que les proportions de l'influenceuse artificielle varient. Et les réponses qu'elle donne dans les chats, entre bonne humeur et soumission, devraient également éveiller les soupçons. Quoi qu'il en soit, le scepticisme est de mise face à la perfection irréaliste et au glamour exagéré.

Les deepfakes à la lumière de différents sondages suisses

Dans l'étude de TA-SWISS sur les deepfakes, plusieurs enquêtes ont été réalisées en plus d'une analyse approfondie de la littérature scientifique, notamment une enquête en ligne, complétée par une expérience en ligne portant sur le vécu et le rapport aux deepfakes de la population. Le groupe de projet a aussi mené plusieurs interviews et sondages auprès de professionnels des médias, du personnel de l'administration et de responsables politiques pour recueillir leur avis sur les risques, les opportunités et les conséquences des trucages vidéo et audio. Enfin, le groupe de projet a testé la capacité de plusieurs détecteurs gratuits de deepfake à détecter les vidéos truquées.

Comment la population et les professionnels des médias perçoivent les deepfakes

En Suisse, la population n'a jusqu'à présent que peu été confrontée aux deepfakes. C'est sur des plateformes comme YouTube, TikTok et Instagram que l'on a le plus de chances de les rencontrer. Les personnes interrogées les associent principalement à des risques et ne sont guère en mesure de distinguer les vidéos deepfakes bien réalisées des vidéos authentiques. Même dans les plus grands médias suisses, les deepfakes sont avant tout perçus comme un risque. Les médias journalistiques ont un rôle important à jouer en termes de sensibilisation de la population à ce thème.

L'étude de TA-SWISS est la première enquête globale à se pencher sur la manière dont les deepfakes sont perçus en Suisse. Sur plus de 1300 personnes interrogées, un peu plus de la moitié a déclaré connaître le terme « deepfake » – et un peu moins de la moitié a déclaré en avoir déjà vu. Dans notre pays, une petite minorité reconnait avoir personnellement testé la création (2%) ou la diffusion (3%) de deepfakes. Les résultats de l'étude de TA-SWISS montrent dans l'ensemble que les gens en Suisse ont plutôt peu d'expérience avec les technologies deepfake. Les paramètres typiques tels que l'âge, le sexe et l'éducation, qui jouent généralement un rôle dans l'acquisition de nouvelles techniques, n'ont pas beaucoup d'influence à cet égard.

La société plus menacée que l'individu

Pour la population suisse, les technologies deepfake comportent plutôt des risques que des opportunités. Les craintes portent en premier lieu sur le fait que les fausses informations diffusées sous forme de deepfakes pourraient affaiblir la confiance dans les médias d'information suisses. Le risque que des deepfakes puissent influencer des votations ou des élections en Suisse est jugé un peu moins virulent.

Pour les personnes interrogées, le risque d'être elles-mêmes victimes de deepfakes est relativement faible. Il est frappant de constater que les femmes estiment que ce danger est plus élevé que les hommes – un résultat guère surprenant au vu des nombreux deepfakes pornographiques.

La perception des opportunités varie selon la dénomination

Lorsqu'on enquête sur les éventuelles opportunités qui pourraient découler des deepfakes, les personnes interrogées se montrent sceptiques. Mais cela change lorsque l'on utilise l'expression plus neutre de « médias synthétiques » au lieu du terme « deepfake ». Dans une étude préliminaire, deux questionnaires différents ont été distribués à deux groupes distincts : l'un d'entre eux employait le terme « deepfake », et l'autre l'expression « médias synthétiques ».

Il s'avère que l'expression « médias synthétiques » est moins connue du public : les personnes interrogées ne sont qu'un peu plus d'un tiers à la connaître, alors qu'environ deux tiers d'entre elles connaissent le terme « deepfakes ». Et, tandis que leurs risques respectifs sont jugés à peu près identiques, ce n'est pas le cas pour leurs opportunités : on attribue aux médias synthétiques nettement plus d'opportunités qu'aux deepfakes en termes d'impact sur les médias et l'économie. La manière dont la société perçoit l'utilité de cette technologie dépend donc en grande partie de sa dénomination.

Inefficacité des astuces, utilité de la familiarisation avec les nouveaux médias

L'étude de TA-SWISS montre à quel point il est difficile de reconnaitre un deepfake. Dans le cadre d'une expérience, les personnes interrogées ont visionné trois vidéos deepfake et trois vidéos authentiques, dont elles devaient évaluer le degré de véracité. Les personnes interrogées ont été réparties en deux groupes, dont l'un a bénéficié au préalable de brèves astuces pour reconnaitre les deepfakes.

En conclusion, les deux groupes se sont révélés très peu sûrs de leur appréciation et guère en mesure de distinguer une vidéo deepfake bien faite d'une vidéo authentique. En outre, le groupe qui avait reçu les conseils préalables pour repérer un deepfake n'a pas mieux su le faire que le groupe qui n'avait pas reçu cette aide.

En revanche, il s'est avéré que l'expérience en matière de médias sociaux était positivement corrélée à la détection de deepfakes. Cela confirme que les compétences médiatiques contribuent à lutter contre les trucages vidéo : il s'agit non seulement de se familiariser avec les médias traditionnels, mais aussi d'apprendre à traiter avec prudence toute information provenant de sources inconnues sur les médias sociaux.

Défis pour le journalisme

Reconnaitre les fausses informations et la désinformation fait partie de la mission principale des professionnels des médias. Les vidéos et les enregistrements audio trompeusement réalistes représentent un défi supplémentaire à cet égard. Comme le journalisme est censé suivre les événements politiques d'un œil critique tout en contribuant à la formation de l'opinion et de la volonté du public, l'identification correcte des deepfakes par les professionnels des médias revêt une importance pour la société dans son ensemble. Mais il est aussi dans l'intérêt des médias de ne pas propager (involontairement) des deepfakes sous peine de gravement entacher leur crédibilité, et donc leur modèle économique.

Les professionnels des médias sont mis au défi de vérifier l'authenticité des vidéos (ou des enregistrements audio) dans un laps de temps aussi court que possible. Mais ces vérifications sont coûteuses. Parallèlement, de nombreux groupes média sont soumis à des pressions financières et ne peuvent pas tous se permettre de recruter du personnel spécialisé dans ce domaine. En outre, si les articles journalistiques peuvent sensibiliser le public aux deepfakes, leur accorder (trop) de place dans les reportages pourrait accroitre le risque de voir se propager un scepticisme excessif au sein de la population et, plus généralement, une méfiance accrue à l'égard des contenus médiatiques.

Les journalistes sont très exposés en public. Comme le montrent des expériences faites en Inde et aux États-Unis, des personnalités médiatiques de premier plan peuvent elles-mêmes être victimes de deepfakes. Et si les journalistes suisses ne sont généralement pas aussi célèbres que certains de leurs collègues à l'étranger, beaucoup redoutent aussi cette menace dans notre pays. Les deepfakes viennent élargir l'arsenal d'intimidation.

Alerte en sourdine dans les salles de rédaction suisses

L'enquête menée auprès des journalistes suisses dans le cadre de l'étude de TA-SWISS montre que le phénomène des deepfakes est certes perçu dans les rédactions et traité dans le cadre de la formation journalistique, mais qu'il n'est pas considéré comme un risque imminent. On estime plutôt que le trucage de vidéos est une sous-catégorie de la désinformation. À l'heure actuelle, les professionnels des médias en Suisse ne craignent pas (encore) d'être eux-mêmes victimes de deepfakes.

Les rédactions sont surtout confrontées aux trucages vidéo dans le cadre de reportages à l'étranger, notamment dans le contexte de la guerre en Ukraine. Dans ce cas, les médias sont appelés à détecter les vidéos truquées afin de ne pas les diffuser à leur insu. À cet égard, selon les résultats de l'enquête, les rédactions locales pourraient bénéficier de la présence d'équipes de recherche bien dotées dans les grands médias étrangers pour vérifier l'authenticité des vidéos. En revanche, il semble que les médias suisses ne se trouvent pas dans le viseur des sites de production de deepfake, car ils jouissent de moins d'attention à l'échelle internationale.

L'enquête menée auprès des médias suisses a également montré que les cas complexes sont très exigeants en matière de vérification – et que des contrôles minutieux dans les rédactions ne suffisent pas, mais doivent être complétés par des explications et une sensibilisation du public. On ne peut pas se contenter d'une analyse critique de la véracité des informations par les médias. Il faut plutôt que la société dans son ensemble prenne conscience du phénomène de manipulation de l'information.

Des sources fiables s'élèvent contre la fronde deepfake

Les professionnels des médias interrogés considèrent majoritairement les deepfakes comme un risque. Tout au plus reconnait-on le potentiel d'une présentation « synthétique » pour personnaliser la diffusion d'actualités, ou d'avatars pour effectuer des recherches.

Le seul avantage que l'on concède aux vidéos et images truquées, c'est de contribuer à renforcer la position des médias journalistiques en tant que source d'information fiable – pour autant que ces derniers parviennent à identifier à temps les deepfakes et autres manipulations et à se démarquer ainsi des sources moins fiables.

Exigences juridiques différentes pour les médias journalistiques et les plateformes en ligne

En ce qui concerne la diffusion de vidéos truquées, les différences entre les dispositions légales applicables aux médias journalistiques et aux plateformes en ligne jouent un rôle important. Les médias traditionnels sont mentionnés dans la Constitution fédérale de la Confédération suisse, dont l'article 93 prescrit que la radio et la télévision doivent présenter les événements de manière fidèle et refléter équitablement la diversité des opinions. La loi fédérale sur la radio et la télévision indique également, parmi les exigences minimales relatives au contenu des programmes, que les faits et les événements doivent être présentés de manière fidèle, « et permettre au public de se faire sa propre opinion. Les vues personnelles et les commentaires doivent être identifiables comme tels. » Il existe également des instances auprès desquelles déposer plainte si les médias ne respectent pas les principes d'équité et d'équilibre.

En revanche, les plateformes en ligne telles que les réseaux sociaux ou les services de partage de vidéos *(video sharing)* qui diffusent le contenu de leurs utilisatrices et utilisateurs ne sont pas tenues d'indiquer si une vidéo est un deepfake. Au contraire, les contenus diffusés sur les médias sociaux sont protégés par la liberté d'expression, de sorte que l'État ne peut agir que contre les contenus manifestement illicites. De plus, en matière de vidéos pornographiques, c'est la loi fédérale sur la protection des mineurs dans les secteurs du film et du jeu vidéo qui s'applique. Cette loi oblige les diffuseurs de tels contenus, y compris les services de streaming, à prendre des mesures pour protéger les mineurs et, le cas échéant, à restreindre l'accès aux vidéos en question. L'application du droit suisse face aux services offerts à l'étranger constitue toutefois l'un des plus grands défis du traitement des deepfakes.

Quand les avatars font de la politique et bousculent l'économie

Que ce soit lors de conflits armés ou de campagnes électorales : quand la situation dégénère, le recours aux deepfakes permet de déstabiliser la partie adverse. Dans le secteur économique, les films truqués s'ajoutent au répertoire de la cybercriminalité.

Plus personne ne s'étonne de voir les deepfakes utilisés dans les campagnes électorales pour discréditer les adversaires politiques et embrouiller le public cible.

Ainsi en mars 2022, un deepfake diffusait un discours du président ukrainien Volodymyr Zelensky appelant la population de son pays à capituler devant les forces russes. Au Pakistan, le leader de l'opposition Imran Khan s'adressait début 2024 à ses compatriotes depuis la prison où il était incarcéré – et intervenait lui-même dans la campagne électorale sous la forme d'un clone généré par une IA. Aux États-Unis également, on a tenté d'influencer la campagne électorale au moyen de différents trucages deepfake. Ainsi, en janvier 2024, certains membres du parti démocrate du New Hampshire ont reçu un faux appel téléphonique de Joe Biden. Dans ce robocall, le président leur demandait de rester à l'écart des primaires.

En Suisse, les médias synthétiques ont aussi déjà fait leur apparition dans des débats politiques. À l'été 2023, une affiche électorale entièrement générée par l'IA et représentant une ambulance entravée par des activistes du climat – un incident qui ne s'est jamais produit sous cette forme dans notre pays – provoquait un tollé. En octobre de la même année, une vidéo truquée montrait la conseillère nationale Sibel Arslan en train de lancer un appel allant totalement à l'encontre de ses valeurs déclarées.

L'humour comme auxiliaire de campagne électorale

C'est un fait : les deepfakes peuvent semer la confusion dans l'opinion publique et discréditer des personnalités politiquement actives, les intimider ou du moins leur soutirer des informations confidentielles. Cependant, malgré leurs effets dévastateurs, les vidéos de synthèse ont également un potentiel positif en politique : les contributions humoristiques sont susceptibles de stimuler le débat politique et d'aider à la formation de l'opinion. Et lors de votations, les deepfakes pourraient servir à illustrer des faits complexes.

Les responsables politiques pourraient tirer parti des deepfakes satiriques et humoristiques –identifiés en toute transparence– pour s'adresser à leur électorat. En effet, l'humour et l'esprit sont des atouts indéniables lorsqu'il s'agit d'élargir la portée d'une action et d'attirer l'attention du public.

Plus de vigilance souhaitée dans le fonctionnement politique

Les autrices et auteurs de l'étude de TA-SWISS ont interrogé des membres du Parlement suisse et du personnel de l'administration fédérale sur leur perception et leur appréciation des deepfakes.

Les vidéos truquées ont fait leur entrée dans le quotidien politique. En effet, une majorité des personnes interrogées ont déclaré que les deepfakes étaient déjà un sujet de préoccupation dans leur travail. À la question de savoir s'il fallait plutôt considérer les trucages vidéo comme un risque ou comme une opportunité, la réponse a été unanime : les personnes interrogées ne leur ont quasiment pas reconnu d'aspects positifs, n'y voyant au contraire que des risques, en particulier pour la démocratie suisse et pour la confiance dans les institutions locales. Le risque d'être elles-mêmes victimes ou protagonistes d'un deepfake, ou qu'une vidéo truquée puisse ternir les relations internationales a également été évoqué – même si de tels événements restent à leur avis plutôt peu probables. De l'avis commun, les mesures de protection concrètes contre les deepfakes sont encore trop rares.

Potentiel pour le divertissement et l'éducation

Les deepfakes se manifestent de manière moins négative dans l'économie que dans la politique et l'administration. L'industrie du divertissement leur reconnait de nombreuses utilités, notamment dans le domaine cinématographique. Dans le milieu du gaming, on cherche des moyens de transférer les

visages des joueurs sur leurs avatars. Et le secteur de la publicité espère tirer profit d'influenceurs artificiels capables de présenter des vêtements ou d'être actifs dans la communication d'entreprise.

Les deepfakes peuvent également aider à collecter des fonds dans le cadre de campagnes menées par des institutions à but non lucratif. En 2019, un sosie de synthèse de l'ancienne star du football David Beckham a appelé, dans neuf langues différentes, à signer une pétition pour la lutte contre le paludisme, demandant ainsi aux leaders des pays les plus touchés par la maladie de s'engager davantage contre celle-ci.

Dans les salles de classe, l'intérêt pour les cours d'histoire pourrait augmenter si des deepfakes de Cléopâtre, Napoléon ou d'autres personnages historiques s'entretenaient de manière interactive avec les jeunes. Leur motivation à apprendre dans le cadre de l'enseignement à distance personnalisé pourrait aussi être renforcée par le recours à des avatars. En médecine également, on voit le potentiel positif des médias synthétiques, notamment pour soigner les troubles de l'anxiété. Cela consiste à placer un avatar de la personne en traitement dans une posture qui, en temps normal, lui fait peur (comme se tenir en équilibre très en hauteur). Sur le plan psychologique, cela permet d'aborder objectivement une situation donnée.

Espionnage économique avec usurpation d'identité

Mais les deepfakes ne sont pas non plus sans risques pour l'économie où ils sont en mesure de causer des dommages de réputation similaires à ceux de la politique : de fausses déclarations privées d'une personne prétendument initiée peuvent ruiner la réputation d'une entreprise ou manipuler les marchés boursiers. Le recours à des influenceurs deepfake est aussi possible dans le cadre de fausses publicités qui finissent par affaiblir la confiance dans un fournisseur.

En outre, les deepfakes facilitent la fraude à l'identité. Une voix clonée ou l'avatar 3D d'un individu permet de déjouer les systèmes de reconnaissance vocale ou faciale. C'est ainsi qu'une organisation criminelle est capable d'accéder au compte personnel d'un particulier ou d'espionner des secrets commerciaux.

En soi, les attaques deepfake contre les acteurs économiques ne soulèvent pas de questions fondamentalement nouvelles, mais elles s'ajoutent au répertoire de la cybercriminalité traditionnelle. En matière d'objectifs, cyberdélinquance et deepfakes se ressemblent également : les intérêts financiers ou le sabotage de la concurrence sont au premier plan. Les entreprises ou personnes qui remplissent les quatre critères suivants sont particulièrement exposées : grande valeur, visibilité élevée, tendance à une certaine inertie dans leurs activités et accès relativement facile.

La Suisse, une cible attrayante

Économiquement parmi les pays les plus innovants et les plus productifs au monde, la Suisse est considérée comme une cible attrayante pour les cyberattaques – et donc les attaques de type deepfake –, d'autant plus que son architecture de sécurité est à la traîne par rapport à son rôle économique majeur : selon le Global Cybersecurity Index 2020, elle se classe en 42e position sur 182 pays. Dans une étude réalisée la même année pour le compte du Service de renseignement de la Confédération, 15% des entreprises interrogées déclaraient avoir déjà été victimes d'espionnage économique. Seule une minorité (13%) des entités concernées ont cependant signalé l'incident à la police ou au procureur. Bien plus souvent, elles ont pris des mesures internes, fréquemment secondées par un soutien externe. Même les fleurons de l'économie du pays ne sont pas à l'abri des cyberattaques, comme l'ont montré en 2023 les attaques contre la Neue Zürcher Zeitung et les CFF.

De nombreuses entreprises restant discrètes et ne signalant pas les cyberattaques, il est impossible de chiffrer les dommages économiques causés par les deepfakes, d'autant plus que seule une partie des attaques en ligne leur est imputable. En revanche, il existe des chiffres sur le marché illégal qui s'est développé autour de ces trucages vidéo : en 2023, pour une vidéo truquée simple, il fallait compter un tarif de vingt dollars américains par minute sur le darkweb. Ces plateformes proposent en complément de leurs services des instructions, des discussions et d'autres aides relatives aux deepfakes. Tandis qu'on trouve principalement des conseils utiles pour la distribution et l'achat de produits et de services liés aux vidéos manipulées sur des forums du darkweb en anglais et en russe, une activité intense à cet égard est aussi enregistrée sur des sites darknet en turc, en espagnol et en chinois.

Les mesures qui permettraient de limiter les risques liés aux deepfakes sont exposées dans le dernier chapitre avec les recommandations de l'étude de TA-SWISS.

Les deepfakes aux yeux de la loi

En Suisse, la législation en vigueur couvre la plupart des infractions qui peuvent être commises avec des deepfakes. Son application se heurte toutefois à des difficultés considérables et requiert une coopération au niveau international.

La liberté artistique, tout comme la liberté d'opinion et d'information, fait partie des libertés fondamentales qui jouissent d'une place essentielle en Suisse. Elles sont garanties à la fois par la Constitution fédérale et par la Convention européenne de sauvegarde des droits de l'homme et des libertés fondamentales. Les deepfakes sont eux aussi protégés par les droits fondamentaux mais cette protection peut être restreinte si un contenu falsifié porte atteinte aux droits d'autrui.

Protection du droit d'auteur pour les créations

Les médias synthétiques bénéficient donc de la protection du droit d'auteur dans la mesure où ils sont considérés comme des œuvres : selon la loi sur le droit d'auteur, cela inclut les productions photographiques, cinématographiques et autres productions visuelles ou audiovisuelles. Le contenu artistique ou esthétique de la vidéo ou de l'image ne joue aucun rôle ; ce qui compte, c'est la contribution créative d'un individu. Or, dans le cas d'une vidéo générée de manière entièrement automatisée par une IA, l'effort créatif fait défaut. On peut donc douter que les vidéos créées sans intervention humaine soient aussi considérées comme des œuvres protégées par le droit d'auteur ou par la liberté artistique.

Limites de la liberté d'information

La liberté d'information a elle aussi ses limites. Les allégations délibérément fausses n'entrent pas dans le champ de cette protection. Pourtant, à l'exception de la loi sur la radio et la télévision, qui impose aux radiodiffuseurs de fournir des informations pertinentes, aucune loi ne prescrit le caractère véridique des vidéos.

En revanche, si des deepfakes menaçants terrorisaient la population, par exemple en l'intimidant avec une catastrophe imminente, leurs autrices et auteurs pourraient être poursuivis en justice. L'article 258 du Code pénal suisse prévoit en effet des peines pour « quiconque jette l'alarme dans la population par la menace ou l'annonce fallacieuse d'un danger pour la vie, la santé ou la propriété ».

Vol d'identité, atteinte à la réputation et fraude par deepfake

Le Code civil suisse protège différents droits de la personnalité – dont le droit à l'image, à la voix et au nom. Tout deepfake à base d'images et d'enregistrements sonores pris sans consentement constitue une violation des droits de la personnalité de la personne en question.

Un individu mis en scène de manière désavantageuse dans un deepfake peut appuyer sa défense sur différents articles du Code pénal suisse qui s'appliquent contre la diffamation et les atteintes à l'honneur. Les dispositions du code pénal peuvent également être invoquées par une personne trompée à des fins frauduleuses – par exemple pour l'inciter à transférer de l'argent ou à révéler des informations confidentielles.

Le droit pénal sanctionne la diffusion de contenus pornographiques – une infraction qui devrait concerner une grande partie des deepfakes produits. Il est illégal d'exposer publiquement des enregistrements sonores ou visuels pornographiques ou de les proposer à quiconque sans y avoir été invité. Si un deepfake est imposé à un individu, le délit de harcèlement sexuel peut également être constitué. Le phénomène du *revenge porn* fait l'objet d'un article récent du code pénal qui punit la transmission non autorisée de contenus sexuels non publics.

Une falsification de documents sophistiquée

Il n'est pas rare que des enregistrements de caméras de surveillance ou de bodycams soient utilisés comme preuves dans des procédures judiciaires. Les technologies deepfake pourraient permettre de truquer ces vidéos ou, à l'inverse, d'en fabriquer pour créer de faux alibis. La falsification de documents n'a rien de nouveau en tant que telle, mais les vidéos et les enregistrements audio générés par une IA leur ajoutent un nouveau chapitre.

Avec les deepfakes, l'appréciation des preuves audiovisuelles se complique encore. Mais quels que soient les moyens utilisés pour manipuler des documents, des vidéos ou des enregistrements sonores relevant du droit procédural, quiconque falsifie un acte officiel commet un délit. Et toute personne qui porte des accusations contre quelqu'un en connaissance de cause peut, à son tour, être poursuivie pour fausses accusations ou tromperie de la justice.

Les médias synthétiques comme outil d'aide à l'application de la loi

Parmi les spécialistes du droit, on débat de l'utilisation des médias synthétiques comme instrument contre le crime. Dans le cadre d'investigations secrètes, il est souvent nécessaire de télécharger du matériel de pornographie enfantine pour infiltrer les forums en ligne concernés. Or, les enquêteurs ne sont pas autorisés à commettre des délits même s'ils traquent des criminels – et dans notre pays, la diffusion de pornographie enfantine « fictive » constitue un délit. Créer un deepfake de pornographie enfantine entièrement synthétique est aussi problématique, car la production d'une vidéo réaliste nécessite des données d'entrainement composées d'images d'abus d'enfants réellement commis. La police n'est donc pas autorisée à utiliser des deepfakes de pornographie enfantine dans le cadre de ses enquêtes.

Enfin, bien qu'il soit envisageable d'utiliser les deepfakes dans le cadre de poursuites pénales – par exemple pour reconstituer le déroulement d'un crime à l'aide de vidéos de téléphones portables, de données de caméras de surveillance et de scanners corporels –, une telle procédure soulève aussi un certain nombre de questions juridiques. Ce qui pose problème, c'est qu'une telle reconstitution, bien qu'elle paraisse objective, repose en fait uniquement sur les hypothèses de l'enquête criminelle. On ne sait pas non plus comment accorder aux prévenus le droit de participer à toutes les étapes de la procédure, y compris la « visite virtuelle de la scène de crime », ni comment les preuves numériques seront versées au dossier.

Coopération internationale dans la lutte contre les agissements mondialisés

La législation suisse couvre donc la plupart des infractions qui peuvent être commises avec des deepfakes.

Mais l'application du droit suisse se heurte souvent à des obstacles importants. En effet, la paternité des deepfakes est généralement difficile à établir. Et même si l'on parvient à identifier ses autrices et auteurs, cela ne sert plus à grand-chose après qu'un deepfake a été diffusé des milliers de fois et s'est dispersé aux quatre vents. La plupart des deepfakes sont d'origine étrangère et sont téléchargés sur des plateformes situées en dehors de la Suisse.

En outre, les infractions commises sur internet entrainent généralement des procédures coûteuses. Il y a souvent plusieurs personnes impliquées, qu'il faut identifier, auxquelles s'ajoutent des responsabilités mal définies et des autorités de poursuite pénale surchargées.

Les spécialistes espèrent que les accords d'entraide judiciaire et le renforcement de la coopération internationale en matière d'échange de données permettront d'améliorer l'application transfrontalière du droit. Dans l'Union européenne, la loi sur les services numériques *(Digital Services Act,* DSA) est un règlement censé assurer une meilleure protection des utilisatrices et utilisateurs sur internet. Les plateformes sont notamment tenues de lutter contre les contenus illégaux. En outre, elles doivent permettre aux utilisatrices et utilisateurs de signaler des contenus et de coopérer avec des *trusted flaggers,* ou signaleurs dits « de confiance ». Il s'agit d'institutions qui traquent les contenus illicites et les signalent à la plateforme. Récemment, l'UE a également adopté un acte législatif sur l'intelligence artificielle qui prévoit une obligation de transparence pour les deepfakes.

Comme les réseaux sociaux, les plateformes en ligne ne sont pas restées inactives. Plusieurs d'entre elles ont élaboré des directives à l'attention de leur communauté pour interdire les trucages numériques et les informations trompeuses. En outre, 34 grands groupes – dont Meta, Google, Microsoft et TikTok – ont signé un code de conduite par lequel ils s'engagent à lutter contre les fausses informations. Mais se contenter de compter sur l'autorégulation des grandes plateformes ne servirait guère l'intérêt public. En effet, les démarches pour définir des critères de suppression manquent de participation démocratique et de transparence. Le risque d'un exercice unilatéral du pouvoir ne doit donc pas être négligé.

Redresser les distorsions de la réalité : quelques recommandations pour gérer les deepfakes

Les conséquences déplaisantes des deepfakes ne peuvent pas être évitées ou freinées uniquement par des mesures réglementaires ou techniques individuelles. Au contraire, pour pouvoir également profiter du potentiel des médias synthétiques, une combinaison de précautions à différents niveaux et beaucoup de responsabilité personnelle sont nécessaires.

La plupart des vidéos manipulées trouvent leur public sur les grandes plateformes en ligne, qui ont donc un rôle clé à jouer dans la régulation des médias synthétiques. Les autorités sont également appelées à agir, tout comme le secteur de la communication, de l'éducation et, bien entendu, chaque citoyenne et citoyen.

Assumer sa responsabilité personnelle

Dans tous les secteurs, la formation aux compétences liées aux médias et à l'information devrait figurer en tête des priorités. Quant à la population, il incombe à chacune et chacun d'utiliser de son propre chef les services d'éducation et de sensibilisation proposés par différents organismes. Faire appel à la responsabilité personnelle est également impératif en matière d'appréciation, de rediffusion et, surtout, de production de deepfakes. En outre, il est important que les gens prennent conscience que le téléchargement d'images et d'enregistrements vocaux peut favoriser la production de médias synthétiques. Le principe selon lequel internet n'oublie pas est tout particulièrement valable pour les deepfakes.

Toute personne qui apprécie les vidéos sur internet, ou en reçoit par le biais des médias sociaux, devrait toujours garder à l'esprit qu'il peut s'agir d'un trucage. Le scepticisme est tout indiqué lorsque le contenu est émotionnellement chargé ou particulièrement spectaculaire.

Responsabiliser les plateformes et renforcer la protection des victimes

L'État devrait s'employer à contraindre les plateformes de supprimer ou de bloquer les deepfakes signalés. De plus, les opérateurs de plateforme devraient être encouragés à mettre en place un système de signalement des deepfakes. Des exigences de transparence et des possibilités de recours renforceraient également les droits des victimes de deepfakes et des personnes concernées par des suppressions injustifiées. Pour mettre en œuvre de telles mesures, la coopération internationale est indispensable et appelle à la création d'instruments de coopération supplémentaires avec d'autres États. En outre, la Suisse devrait s'engager pour que des normes et des règles contre la cybercriminalité soient définies et appliquées au niveau international.

En cas de conflit avec de grandes plateformes en ligne, les particuliers sont généralement perdants. C'est pourquoi il faut des services spécialisés qui conseillent et soutiennent les victimes de deepfakes – ou les personnes concernées par des suppressions injustifiées. La Confédération et les cantons devraient doter les services d'aide aux victimes de cyberdélits de moyens humains et financiers suffisants. Il faudrait aussi que la Suisse reconnaisse des trusted flaggers, de sorte que la priorité soit accordée à leurs signalements de deepfakes problématiques sur internet, et qu'elle envisage même un soutien financier pour ces signaleurs de confiance.

Le progrès technique pour se défendre

Il faut encourager un débat de fond sur les procédures d'authentification et d'identification. Certaines méthodes avancées, en particulier l'authentification multifactorielle, peuvent aider à déjouer les tentatives de tromperie par deepfake vocal ou facial. Dans la mesure du possible, il est important que les méthodes d'authentification les plus sophistiquées soient employées, car la cybercriminalité s'emploie de son côté à contourner les mesures de protection.

Compte tenu de l'évolution rapide des technologies deepfake, il faut recourir à tous les instruments possibles pour empêcher les abus. Même des outils encore peu efficaces à l'heure actuelle, comme les détecteurs de deepfake, constituent potentiellement un élément de mosaïque d'une protection globale. Enfin, il est recommandé de renforcer au maximum les mesures de sécurité existantes.

Sensibiliser aux risques – et aux avantages

Aujourd'hui, peu de gens ont une expérience personnelle des deepfakes, et beaucoup n'en savent pas grand-chose. L'information dans les salles de classe et dans les médias devrait sensibiliser au phénomène, notamment en fournissant des conseils sur la manière de vérifier les sources et de remettre en question la plausibilité des vidéos. Les écoles devraient vérifier si l'étude des deepfakes pourrait faire partie des objectifs des plans d'études romand (PER) et alémanique (LP21) visant à renforcer les compétences médiatiques.

Malgré les risques auxquels il est nécessaire de sensibiliser, il ne faut pas étouffer le potentiel des vidéos de synthèse. Il convient donc d'être prudent dans le choix de la formulation. En effet, les gens associent beaucoup moins le terme « deepfake » à des opportunités qu'une expression plus neutre comme « média synthétique ».

Membres du groupe d'accompagnement

- **Prof. Reinhard Riedl**, Haute école spécialisée bernoise, président du groupe d'accompagnement, membre du comité directeur de TA-SWISS
- **Dr Bruno Baeriswyl**, expert en protection des données, président du comité directeur de TA-SWISS
- **Cornelia Diethelm**, Centre for Digital Responsibility
- **Prof. Rainer Greifeneder**, directeur du département de psychologie sociale, Université de Bâle
- **Thomas Häussler**, division Médias, Office fédéral de la communication OFCOM
- **Andrea Hauser**, informaticienne, experte en cybersécurité, société de sécurité Scip
- **Erich Herzog**, avocat, membre de la direction d'Economiesuisse
- **Prof. Selina Ingold**, IDEE Institut für Innovation, Design & Engineering, Ostschweizer Fachhochschule
- **Melanie Kömle Bender**, documentaliste médias, SRF Schweizer Radio und Fernsehen
- **Thomas Müller**, journaliste scientifique, membre du comité directeur de TA-SWISS
- **Prof. René Schumann**, HES-SO Valais-Wallis, Institut de recherche en informatique
- **Prof. Giatgen Spinas**, Université de Zurich, membre du comité directeur de TA-SWISS
- **Dr Stefan Vannoni**, économiste, CEO cemsuisse, membre du comité directeur de TA-SWISS

Gestion du projet chez TA-SWISS

- **Dre Elisabeth Ehrensperger**, directrice
- **Dre Laetitia Ramelet**, responsable de projet
- **Dre Lucienne Rey**, responsable de projet

Impressum

Nos yeux et nos oreilles à l'épreuve
Synthèse de l'étude « Deepfakes et réalités manipulées »
TA-SWISS, Berne 2024
TA 81A/2024

Rédaction : Lucienne Rey
Traduction : Alexandra de Bourbon, pro-verbial sàrl, Zurich
Production : Laetitia Ramelet et Fabian Schluep, TA-SWISS, Berne
Mise en page et illustrations : Hannes Saxer, Berne
Impression : Jordi AG – Das Medienhaus, Belp

TA-SWISS – Fondation pour l'évaluation des choix technologiques

Souvent susceptibles d'avoir une influence décisive sur la qualité de vie des gens, les nouvelles technologies peuvent en même temps comporter des risques nouveaux, qu'il est parfois difficile de percevoir d'emblée. La Fondation pour l'évaluation des choix technologiques TA-SWISS s'intéresse aux avantages et aux risques potentiels des nouvelles technologies qui se développent dans les domaines « biotechnologie et médecine », « numérisation et société » et « énergie et environnement ». Ses études s'adressent tant aux décideurs du monde politique et économique qu'à l'opinion publique. TA-SWISS s'attache, en outre, à favoriser par des méthodes participatives, l'échange d'informations et d'opinions entre les spécialistes du monde scientifique, économique et politique et la population. TA-SWISS se doit, dans ses projets sur les avantages et les risques potentiels des nouvelles technologies, de fournir des informations aussi factuelles, indépendantes et étayées que possible. Elle y parvient en mettant chaque fois sur pied un groupe d'accompagnement composé d'experts choisis de manière à ce que leurs compétences respectives couvrent ensemble la plupart des aspects du sujet à traiter.

La fondation TA-SWISS est un centre de compétence des Académies suisses des sciences.

TA-SWISS
Fondation pour l'évaluation
des choix technologiques
Brunngasse 36
CH-3011 Berne
info@ta-swiss.ch
www.ta-swiss.ch

membre des
académies suisses
des sciences