7-392/1

SÉNAT DE BELGIQUE

SESSION DE 2022-2023

9 NOVEMBRE 2022

Proposition de résolution visant à prendre des mesures pour lutter contre les *deepfakes*

(Déposée par M. Tom Ongena et consorts)

DÉVELOPPEMENTS

I. INTRODUCTION: *DEEPFAKES*

A. Contexte

Le terme «*deepfake*» désigne la forme la plus moderne qui soit de manipulation des supports visuels. La manipulation d'images ne date pas d'hier – songeons aux photos truquées –, mais elle atteint un *summum* avec les *deepfakes*. Les conséquences peuvent être lourdes, tant au niveau personnel qu'au niveau sociétal ou géopolitique.

Le *deepfake*, ou hypertrucage, est une technique de création d'images, de sons ou de textes à l'aide de logiciels d'intelligence artificielle. Elle permet d'intervenir sur des images mobiles, en manipulant des visages par exemple, à un point tel que l'on peut avoir l'illusion qu'il s'agit d'un enregistrement d'une personne alors que ce n'est pas le cas. À l'œil nu, il est presque impossible de faire la distinction entre de telles images trompeuses et des images authentiques, ce qui induit un risque majeur d'abus.

«*Deepfake*» est un mot-valise composé à partir des termes *deep learning* («apprentissage en profondeur») et *fake* («faux», «contrefait»). L'apprentissage en profondeur est un apprentissage automatique (soit le fait pour un ordinateur d'apprendre par lui-même) qui utilise l'intelligence artificielle (IA), généralement

BELGISCHE SENAAT

ZITTING 2022-2023

9 NOVEMBER 2022

Voorstel van resolutie om maatregelen te nemen tegen *deepfakes*

(Ingediend door de heer Tom Ongena c.s.)

TOELICHTING

I. INLEIDING: *DEEPFAKES*

A. Probleemstelling

Deepfakes zijn de nieuwste vorm van gemanipuleerd beeldmateriaal. Hoewel beeldmateriaal al heel lang gemanipuleerd wordt, denk aan getrukeerde foto's, zijn *deepfakes* de spreekwoordelijke overtreffende trap. Dit heeft mogelijk grote gevolgen op zowel persoonlijk, maatschappelijk als geopolitiek vlak.

Deepfakes zijn beelden, geluiden of teksten die door slimme software worden gecreëerd. Deze techniek maakt het mogelijk om bewegende beelden zodanig te bewerken, door bijvoorbeeld de gezichten te manipuleren, dat het lijkt alsof van iemand beelden zijn opgenomen terwijl dat niet zo is. Met het blote oog zijn dergelijke bedrieglijke beelden amper te onderscheiden van «echte» beelden, wat het risico op misbruik heel groot maakt.

Het woord «*deepfake*» is een samentrekking van de woorden «*deep learning*» en «*fake*» (vals). Met deep learning bedoelt men machinaal leren (computers die zichzelf iets aanleren) waarbij men gebruik maakt van artificiële intelligentie (AI), vaak via neurale netwerken. Met andere woorden, hoe meer men een netwerk laat

par le biais de réseaux neuronaux. En d'autres termes, plus un réseau est «entraîné» à faire de la manipulation d'images, plus il est performant à cet égard.

Si le terme lui-même n'est connu que depuis quelques années, l'origine de la technologie visée remonte à la fin des années 1990. En 1997, Christoph Bregler, Michele Covell et Malcolm Slaney ont développé un programme véritablement unique et novateur pour l'époque, qui automatisait un processus ne pouvant alors être réalisé que par certains studios cinématographiques. Ce programme, intitulé «*Video Rewrite*», pouvait réaliser de nouvelles animations faciales à partir d'une piste audio. En d'autres termes, il pouvait synchroniser parfaitement des lèvres en mouvement avec le contenu prononcé (1).

Le programme s'appuyait sur des avancées technologiques antérieures, qui permettaient d'interpréter des visages, de créer un audio de synthèse à partir d'un texte et de modeler des lèvres dans un environnement 3D, mais il a été le premier programme à réunir tous ces processus et à permettre la réalisation d'une animation convaincante.

La technologie s'est affinée au fil des décennies suivantes. En 2016 et en 2017, deux publications ont apporté la preuve que des *deepfakes* pouvaient être réalisés avec du matériel informatique disponible sur le marché. Il s'agit du projet *Face2Face* de l'Université technique de Munich et du projet *Synthesizing Obama* de l'Université de Washington. Les deux projets, qui poursuivaient des objectifs totalement différents, ont réussi à améliorer radicalement les temps de calcul et de restitution tout en perfectionnant la fidélité graphique avec un réalisme photographique.

L'essor exponentiel des *deepfakes* est en grande partie imputable à la diffusion de vidéos pornographiques sur la plateforme *Reddit,* un phénomène mis en lumière par la journaliste Samantha Cole du magazine *Vice*. Le sous-reddit *r/deepfakes*, qui a entre-temps été interdit, comptabilisait pres de 90 000 membres et contenait des séquences de pornographie truquée mettant en scène une grande diversité d'acteurs. Après cette interdiction, *Reddit* a mis à jour sa politique de contenu pour mieux faire respecter sa vision concernant les contenus pornographiques.

B. *Deepfakes* visuels

Les exemples les plus célèbres de *deepfake* sont ceux où le visage d'une personne est superposé sur celui

(1) *https://medium.com/@songda/a-short-history-of-deepfakes-604ac7be6016.*

«trainen» op het manipuleren van beelden, hoe beter het hierin wordt.

Het woord is pas sinds enkele jaren bekend, maar de eigenlijke origine van deze technologie ligt in de late jaren '90. In 1997 ontwikkelden Christoph Bregler, Michele Covell en Malcolm Slaney een voor die tijd innovatief, werkelijk uniek programma dat in wezen automatiseerde wat enkel sommige filmstudio's konden doen. Dit programma, genaamd «*Video Rewrite Program*» kon nieuwe gezichtsanimaties samenstellen vanuit een audio output. Met andere woorden, het kon bewegende lippen perfect laten aansluiten op datgene dat gesproken werd (1).

Het programma bouwde voort op ouder werk dat gezichten interpreteerde, audio synthetiseerde uit tekst, en lippen modelleerde in een 3D-ruimte maar het was het eerste waarbij dit alles werd samengevoegd en op een overtuigende wijze geanimeerd.

Deze technologie werd in de decennia die daarop volgden verder verfijnd. In 2016 en 2017 werd in twee publicaties aangetoond dat *deepfakes* haalbaar zijn met consumentenhardware: het *Face2Face*-project van de Technische Universiteit van München en het *Synthesizing Obama*-project van de Universiteit van Washington. Hoewel ze totaal verschillende doelen nastreefden, verbeterden ze drastisch de reken- en renderingtijden, terwijl ze de grafische getrouwheid bijwerkten op een manier die er fotorealistisch uitzag.

De enorme toename van *deepfakes* kan grotendeels worden toegeschreven aan *Reddit* en pornografie, onder de aandacht gebracht door Samantha Cole van *Vice* magazine. Een nu verwijderde *subreddit* met de toepasselijke naam *r/deepfakes* had bijna 90 000 leden en bevatte *deepfake* porno van een verscheidenheid aan acteurs. Na het verbod heeft *Reddit* zijn inhoudsbeleid bijgewerkt om hun standpunt over pornografie beter te handhaven.

B. Visuele *deepfakes*

Het bekendste voorbeeld van *deepfakes* zijn die waarbij men het gezicht van een persoon als het ware

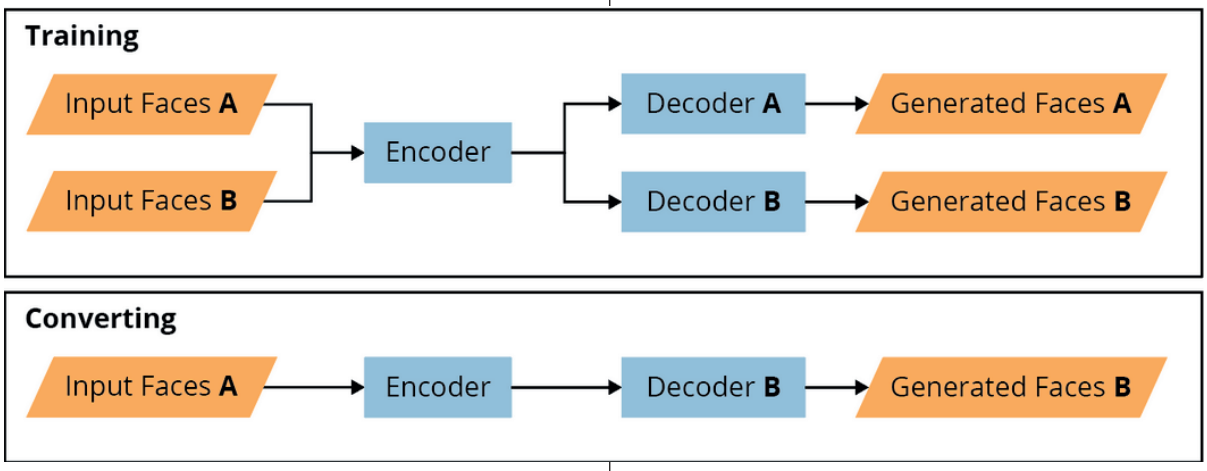
(1) *https://medium.com/@songda/a-short-history-of-deepfakes-604ac7be6016.*

d'une autre. La technique est utilisée aussi bien à des fins ludiques (divertissement, art, etc.) qu'à des fins qui ne le sont guère (vengeance pornographique, fraude, désinformation, etc.). Il existe toutefois différents types de manipulation faciale, allant du *morphing* (création d'un nouveau visage par fusion de deux visages) au *face generation* (création de nouveaux visages inexistants réalistes et projection de ceux-ci sur des visages existants).

Actuellement, *DeepFaceLab* et *FaceSwap* sont les deux infrastructures logicielles *open source* les plus courantes pour la création de *deepfakes*. Leur utilisation, gratuite et libre, est soutenue par de grandes communautés en ligne comptant des milliers d'utilisateurs, dont beaucoup participent activement à l'évolution et à l'amélioration des logiciels et des modèles. Grâce à ce développement permanent, la création de *deepfakes* sera de plus en plus aisée pour des utilisateurs moins avancés, avec à la clé une fidélité accrue et un plus grand potentiel de création de faux contenus médiatiques crédibles (2).

«projecteert» op dat van een ander. Dit wordt meestal voor zowel ludieke (entertainment, kunst, enz.) als minder ludieke doeleinden (wraakporno, fraude, desinformatie, enz.) gebruikt. Er bestaan echter verschillende soorten gezichtsaanpassingen, gaande van morphen (twee gezichten «mixen» tot een nieuw gezicht) tot «*face generation*» (het genereren en projecteren van realistische maar nieuwe, onbestaande gezichten op bestaande gezichten).

Op dit ogenblik zijn *DeepFaceLab* en *FaceSwap* de twee meest gebruikte *open-source softwareframeworks* voor het maken van *deepfakes*. Zij zijn vrij te gebruiken en *open source* en worden ondersteund door grote en toegewijde online gemeenschappen met duizenden gebruikers, van wie velen actief deelnemen aan de evolutie en verbetering van de software en de modellen. Door deze voortdurende ontwikkeling zullen *deepfakes* steeds gemakkelijker te maken zijn voor minder geavanceerde gebruikers, met een grotere getrouwheid en een groter potentieel om geloofwaardige «valse» media te creëren (2)).



Le processus de création d'une vidéo *deepfake*, aussi appelée «infox vidéo», comporte généralement cinq phases, allant de l'identification des visages sur le matériel visuel fourni aux processus de post-édition visant à supprimer les petites erreurs sur le résultat obtenu, en passant par la phase d'entraînement et la conversion des visages. À l'heure actuelle, ces processus sont en grande partie automatisés grâce à l'omniprésence d'applications modernes de *deepfake*.

C. *Deepfakes* audio

La technologie de clonage vocal permet à des ordinateurs de créer une imitation de voix humaine. Les technologies de clonage vocal sont également connues sous le nom de *deepfakes* audio, synthèse vocale ou conversion

Over het algemeen bestaat het proces van het maken van een *deepfake* uit vijf stappen, gaande van het identificeren van de gezichten op het aangereikte beeldmateriaal, het trainen en converteren van de gezichten tot het postprocessen van het reeds verkregen resultaat om kleine fouten weg te werken. Met de alomtegenwoordigheid van moderne *deepfake apps* zijn dergelijke processen tegenwoordig grotendeels geautomatiseerd.

C. Audio-*deepfakes*

De technologie van het klonen van stemmen stelt computers in staat een imitatie van een menselijke stem te creëren. Technologieën voor het klonen van stemmen zijn ook bekend als audio-grafische *deepfakes*

(2) *https://insights.sei.cmu.edu/blog/how-easy-is-it-to-make-and-detect-a-deepfake/.*

(2) *https://insights.sei.cmu.edu/blog/how-easy-is-it-to-make-and-detect-a-deepfake/.*

vocale*/swapping*. Les méthodes logicielles de clonage vocal permettent de générer une voix de synthèse qui ressemble étonnamment à une véritable voix humaine. Certains estiment que la distinction entre une vraie voix et une voix de synthèse «devient imperceptible pour la personne lambda» (traduction) (3).

Le développement de logiciels permettant de cloner des voix par intelligence artificielle a commencé il y a plusieurs décennies, lorsqu'ont été inventées une série de méthodes permettant à des ordinateurs de synthétiser des voix. Les algorithmes de conversion de texte par synthèse vocale (*Text-to-Speech* – TTS) peuvent convertir un texte en mots parlés, ce qui a permis à des ordinateurs d'interagir vocalement avec des humains. Dans de nombreux cas, par exemple les systèmes d'annonce dans les gares, les messages audio traditionnels sont remplacés par un système TTS, qui offre une flexibilité beaucoup plus grande car il n'est plus nécessaire d'enregistrer préalablement tous les messages possibles et imaginables.

La qualité des résultats obtenus au moyen de systèmes TTS basés sur l'intelligence artificielle ne cesse de s'améliorer. Les modèles sont actuellement capables d'apprendre en identifiant de nouvelles structures dans les données audio. L'invention de réseaux antagonistes génératifs (*generative adversarial networks* – GAN), à savoir des algorithmes d'apprentissage automatique capables d'analyser une série d'images donnée et de créer de nouvelles images de qualité équivalente, a non seulement joué un rôle-clé dans la prolifération de *deepfakes* graphiques, mais elle a également joué un rôle de catalyseur dans le développement du clonage vocal, avec à la clé des clones vocaux de plus en plus convaincants et plus difficilement détectables par l'oreille humaine.

L'utilisation de technologies d'intelligence artificielle apporte donc une nouvelle dimension à la crédibilité des clones et à la rapidité avec laquelle il est possible de créer un clone crédible. Le clone n'est cependant pas uniquement convaincant en raison des caractéristiques sonores de la voix. Le contenu du fragment audio doit également correspondre au style et au vocabulaire usuel de la cible.

II. *DEEPFAKES* ET UTILISATION ABUSIVE

A. Utilisation abusive de nature physique

Un rapport publié en octobre 2019 par la *start-up* néerlandaise *Deeptrace*, spécialiste de la cybersécurité,

(3) *https://www.europarl.europa.eu/RegData/etudes/STUD/2021/690039/EPRS_STU(2021)690039_EN.pdf.*

(audio-*deepfakes*), spraaksynthese of *voice conversion/swapping*. Softwaremethoden voor het klonen van stemmen kunnen synthetische spraak genereren die opmerkelijk veel lijkt op een echte, natuurlijke menselijke stem. Sommigen menen dat het verschil tussen een echte en een gesynthetiseerde stem «onmerkbaar wordt voor de gemiddelde persoon» (3).

De ontwikkeling van software voor het klonen van AI-stemmen begon tientallen jaren geleden, toen een aantal methoden werd uitgevonden waarmee computers stemmen konden synthetiseren. Deze zogenaamde *Text-to-Speech* (TTS) algoritmen zijn in staat om tekst om te zetten in gesproken woorden. Hierdoor konden computers spraak gebruiken voor interactie met mensen. In veel gevallen – zoals aankondigingssystemen in treinstations – zijn traditionele audioboodschappen vervangen door een TTS-systeem, waardoor niet langer elk mogelijk bericht vooraf hoeft te worden opgenomen en de flexibiliteit veel groter wordt.

De kwaliteit van de output van op AI gebaseerde TTS-systemen wordt steeds beter. Tegenwoordig zijn de modellen in staat te leren op basis van de ontdekking van nieuwe patronen in audiogegevens. De uitvinding van GAN's – die ook een sleutelrol spelen bij de versnelde ontwikkeling van grafische *deepfakes* (*Generatieve Adversariale Netwerken* zijn algoritmen voor machinaal leren die een gegeven reeks beelden kunnen analyseren en nieuwe beelden met een vergelijkbaar kwaliteitsniveau kunnen creëren) – heeft ook de ontwikkeling van stemklonen versneld, met als resultaat steeds overtuigendere klonen die moeilijker te detecteren zijn door mensen.

Het gebruik van AI-technologie geeft dus een nieuwe dimensie aan de geloofwaardigheid van klonen en aan de snelheid waarmee een geloofwaardige kloon kan worden gecreëerd. Het is echter niet alleen het geluid van een stem dat van een kloon een overtuigende kloon maakt. De inhoud van het audiofragment moet ook overeenkomen met de stijl en de woordenschat van het doelwit.

II. *DEEPFAKES* EN MISBRUIK

A. Fysiek misbruik

Een rapport dat in oktober 2019 werd gepubliceerd door de Nederlandse cyberbeveiligingsstartup *Deeptrace*

(3) *https://www.europarl.europa.eu/RegData/etudes/STUD/2021/690039/EPRS_STU(2021)690039_EN.pdf.*

estimait que 96 % de tous les *deepfakes* publiés en ligne étaient pornographiques (4).

C'est souvent dans des applications pornographiques que les innovations numériques sont largement utilisées au départ. Les *deepfakes* consistent à coller le visage de célébrités sur celui d'acteurs et d'actrices pornos en pleine activité. Ce type de *deepfakes* est parfois appelé «*deepnudes*».

La toute grande majorité de ces *deepfakes*, soit environ 90 %, prennent des femmes pour cibles. En 2020, *Sensity AI* a révélé qu'un *chatbot* (dialogueur) de la messagerie *Telegram* pouvait servir à fabriquer de faux nus (*deepnudes*). La seule chose que les utilisateurs devaient faire était de fournir au *chatbot* la photographie d'une personne qui était ensuite «déshabillée» par des procédés numériques. Le *bot* produisait exclusivement des nus de femmes. Avant que cette information ne soit révélée, plus de 100 000 femmes en avaient déjà été victimes (5).

Ces statistiques mettent en lumière un problème fondamental: le caractère genré des abus et de l'exploitation liés aux *deepfakes*.

Des images de la présentatrice néerlandaise Dionne Stax ont ainsi été truquées et elle a été présentée dans une vidéo à caractère sexuel sur le site pornographique *Pornhub* (6).

Sur le réseau social *Reddit*, les principales rubriques (*subreddits*) consacrées à des *deepfakes* ont été supprimées, hormis les applications inoffensives de la technologie qui avaient souvent une portée ludique. *Twitter* et *Pornhub* s'emploient eux aussi activement à éliminer les vidéos ayant explicitement recours à l'IA. La raison est toujours la même: les vidéos ont été mises en ligne sans le consentement de la célébrité qui y est présentée. Sur toutes les plateformes, il s'agit d'une violation des règles d'utilisation. *Pornhub* place les *deepfakes* dans la même catégorie que la vengeance pornographique. «Nous n'acceptons aucun contenu publié sur le site web sans consentement et supprimons tous les contenus de ce genre dès que nous en sommes informés», déclare un porte-parole de *Pornhub* (7).

Il n'est pas étonnant que les *deepfakes* aillent de pair avec la vengeance pornographique. Précédemment, il

(4) *https://regmedia.co.uk/2019/10/08/deepfake_report.pdf.*
(5) *https://www.europarl.europa.eu/RegData/etudes/STUD/2021/690039/EPRS_STU(2021)690039_EN.pdf.*
(6) *https://www.rtlnieuws.nl/tech/artikel/5199196/deepnudes-deepfakes-manipulatie-fotos-internet.*
(7) *https://techpulse.be/nieuws/222753/reddit-twitter-en-pornhub-verbieden-nep-porno/.*

schatte dat 96 % van alle *deepfakes* online pornografisch was (4).

Zoals bij veel digitale innovatie, vindt er vaak een initiële boom plaats in pornografische toepassingen hiervan. *Deepfakes* worden aangewend om gezichten van beroemdheden op pornoacteurs en -actrices te plakken tijdens hun werkzaamheden. Dit soort *deepfakes* wordt soms ook «*deepnudes*» genoemd.

De overgrote meerderheid van die *deepfakes*, zo'n 90 %, is gericht op vrouwen. In 2020 berichtte *Sensity AI* over een *Telegram chatbot* die kan worden gebruikt om *deepnude* portretten te maken. Het enige wat gebruikers hoefden te doen, was de *chatbot* voorzien van een foto van iemand, die vervolgens digitaal werd «uitgekleed». De *bot* maakte uitsluitend naaktfoto's van vrouwen. Tegen de tijd dat het nieuws openbaar werd gemaakt, waren al meer dan 100 000 vrouwen hiervan het slachtoffer geworden (5).

Deze statistieken brengen een belangrijk probleem aan het licht: de genderspecifieke impact van *deepfake* misbruik en uitbuiting.

Zo werden beelden van de Nederlandse presentatrice Dionne Stax misbruikt om het te doen laten lijken alsof ze in een seksueel getinte video speelde op de pornowebsite *Pornhub* (6).

Op *Reddit* werden de belangrijkste *subreddits* gewijd aan *deepfakes* verwijderd, met uitzondering van onschuldige toepassingen van de technologie, die vaak ludiek bedoeld zijn. Ook op *Twitter* en *Pornhub* worden expliciete AI-video's actief geweerd. De reden is altijd dezelfde: de video's zijn online geplaatst zonder de toestemming van de beroemdheid die erin figureert. Dat gaat bij alle platformen in tegen de gebruiksregels. *Pornhub* deelt de *deepfakes* in bij dezelfde categorie als wraakporno. «We tolereren geen non-consensuele content op de website en verwijderen alle dergelijke content zodra we er op de hoogte van worden gebracht», aldus een woordvoerder van *Pornhub* (7).

Dat *deepfakes* en wraakporno samenvallen hoeft niet te verbazen. Gefotoshopte hoofden van geviseerde

(4) *https://regmedia.co.uk/2019/10/08/deepfake_report.pdf.*
(5) *https://www.europarl.europa.eu/RegData/etudes/STUD/2021/690039/EPRS_STU(2021)690039_EN.pdf.*
(6) *https://www.rtlnieuws.nl/tech/artikel/5199196/deepnudes-deepfakes-manipulatie-fotos-internet.*
(7) *https://techpulse.be/nieuws/222753/reddit-twitter-en-pornhub-verbieden-nep-porno/.*

était déjà arrivé que le visage photoshopé de personnes prises pour cibles soit placé sur du matériel pornographique dans le but de nuire à ces personnes. Il est toutefois beaucoup plus difficile de remarquer à l'œil nu que les *deepfakes* sont des images «manipulées» et de nombreuses personnes attribuent encore aux images animées une «autorité absolue» quant à leur authenticité, étant donné que, jusqu'ici, il était quasi impossible de manipuler de telles images en dehors des grands studios cinématographiques spécialisés dans ce domaine.

Comme les *deepfakes* et *deepnudes*, la vengeance pornographique a été maintes fois abordé ces derniers temps à la Chambre des représentants (8). Il s'agit en l'occurrence de formes nouvelles de voyeurisme et de l'utilisation de l'image de certaines personnes sans leur consentement. Le problème est encore plus grave lorsque ce sont des images de mineurs qui sont ainsi manipulées.

B. Préjudice sociétal et (inter)national

Dans le milieu politique, la diffusion d'images truquées dans le but de nuire au camp opposé n'est pas un phénomène nouveau. Il s'agit souvent de mauvaises images photoshopées dont la manipulation se remarque quasi instantanément. Les *deepfakes* sont plus dangereux car leur manipulation ne saute pas aux yeux.

Jusqu'à présent, c'est la vidéo truquée de Sophie Wilmès, alors première ministre, qui est l'exemple le plus connu d'hypertrucage dans notre pays. On la voyait prononcer un discours fictif dans lequel elle reconnaissait que la pandémie avait des causes identiques à celles de la crise écologique. Elle n'avait en réalité jamais prononcé ces paroles. Il s'agissait d'une initiative du groupe d'activistes du climat *Extinction Rebellion*. Grâce à l'intelligence artificielle, les propos de Sophie Wilmès dans la vidéo truquée paraissaient réels car on entendait et voyait réellement la première ministre prononcer ces mots (9).

Le rapport d'information du Sénat concernant la nécessaire collaboration entre l'autorité fédérale et les Communautés en matière de lutte contre les infox (*fake news*), qui date de 2021, s'inquiétait lui aussi déjà des *deepfakes* et de leur potentiel de désinformation et de déstabilisation (10).

(8) *http://www.dekamer.be/doc/flwb/pdf/55/2141/55k2141006.pdf.*
(9) *https://www.nieuwsblad.be/cnt/dmf20200414_04921988*.
(10) *https://www.senate.be/informatieverslagen/7-110/Senat_rapport_fake_news-2021.pdf.*

personen op pornografisch materiaal werden eerder al gebruikt om schade toe te brengen. *Deepfakes* echter zijn veel moeilijker met het blote oog te beoordelen als «gemanipuleerd» en voor velen hebben bewegende beelden nog steeds een «absolute autoriteit» op het gebied van authenticiteit, aangezien tot nog toe bewegende beelden amper te manipuleren vielen buiten de grote filmstudio's die zich daarmee bezig houden.

Wraakporno in het bijzonder werd recentelijk reeds verschillende keren in de Kamer van volksvertegenwoordigers aangehaald, samen met *deepfakes* en *deepnudes* (8). Het betreft dan ook nieuwe vormen van voyeurisme en het gebruik van de beeltenis van personen zonder hun toestemming. Problematischer wordt het als dergelijke manipulatie plaatsvindt op beelden waarbij minderjarigen gebruikt worden.

B. Maatschappelijke en (inter)nationale schade

Het is niet nieuw dat in de politiek gemanipuleerde beelden worden verspreid om het tegengestelde kamp schade toe te brengen. Het gaat dan vaak om slechte *photoshops* waarbij de manipulatie nagenoeg onmiddellijk zichtbaar is. *Deepfakes* zijn gevaarlijker omdat dit niet onmiddellijk zichtbaar is.

Op dit ogenblik is de *deepfake* video van toenmalig eerste minister Sophie Wilmès het bekendste voorbeeld in ons land. Daarin hield ze een fictieve speech waarin ze erkende dat de pandemie dezelfde oorzaken had als de ecologische crisis. Zelf had ze die woorden nooit uitgesproken. Het was het initiatief van de klimaatactiegroep *Extinction Rebellion*. Het *deepfake*-filmpje laat de speech van Sophie Wilmès er dankzij artificiële intelligentie zo levensecht mogelijk uitzien, want je hoorde en zag de eerste minister de woorden precies echt uitspreken (9).

In het informatieverslag van de Senaat over de noodzakelijke samenwerking tussen de federale overheid en de Gemeenschappen inzake de bestrijding van *fake news* van 2021 werden ook reeds zorgen geuit over *deepfakes* en het potentieel hiervan voor desinformatie en destabilisatie (10).

(8) *http://www.dekamer.be/doc/flwb/pdf/55/2141/55k2141006.pdf.*
(9) *https://www.nieuwsblad.be/cnt/dmf20200414_04921988*.
(10) *https://www.senate.be/informatieverslagen/7-110/Senaat_verslag_fake_news-2021.pdf.*

«Les phénomènes tels que la désinformation et le trucage de vidéos prennent de plus en plus d'ampleur dans les médias mais également dans la vie politique. Il s'agit d'un phénomène tant national qu'international. Ces procédés engendrent de grands risques de radicalisation, de polarisation et de manipulation par des autorités étrangères ou par des groupements extrémistes qui menacent notre société.

Leur objectif est clair: par l'utilisation systématique de la désinformation, ils veulent semer la discorde, miner la confiance dans nos institutions démocratiques et ainsi, déstabiliser et affaiblir notre société tout entière», précise le rapport.

Les *deepfakes* qui peuvent manifestement s'appliquer en temps réel vont encore plus loin. Ainsi, en juin 2022, les maires de plusieurs capitales européennes ont été piégés par un *deepfake* et amenés à s'entretenir en visioconférence avec quelqu'un se faisant passer pour leur homologue de Kiev, Vitali Klitschko.

La maire de Berlin, Franziska Giffey, a participé à un entretien planifié, sur la plateforme de visioconférence *Webex*, avec une personne qui, selon elle, avait l'apparence et la voix de M. Klitschko (11).

«Rien n'indiquait que l'entretien en visioconférence ne se déroulait pas avec une personne réelle», a-t-elle déclaré.

Ce n'est qu'au bout d'environ quinze minutes, lorsque le prétendu maire de Kiev, avec qui la liaison avait été établie, s'est mis à parler du problème des réfugiés ukrainiens qui abusent des allocations sociales allemandes et a semblé en appeler au renvoi des réfugiés en Ukraine pour qu'ils puissent y effectuer un service militaire que Mme Giffey a commencé à avoir des doutes.

Lors d'une brève interruption de la liaison, le bureau de la maire de Berlin a contacté l'ambassade d'Ukraine en Allemagne, laquelle a confirmé, renseignements pris auprès des autorités de Kiev, que la personne qui prenait part à la visioconférence n'était pas le véritable M. Klitschko, a indiqué le quotidien allemand *Der Spiegel*.

Indépendamment du fait que les *deepfakes* sont utilisés pour faire du tort, le fait que l'on se mette à douter de tout est encore plus préjudiciable à la démocratie et

(11) *https://www.theguardian.com/world/2022/jun/25/european-leaders-deepfake-video-calls-mayor-of-kyiv-vitali-klitschko.*

«Verschijnselen als desinformatie en *deepfakes* zijn steeds nadrukkelijker aanwezig in de media, maar ook in de politiek, en dat op nationaal en internationaal niveau. Hierin schuilen grote gevaren voor radicalisering, polarisering en beïnvloeding door zowel buitenlandse overheden als extremistische groeperingen die een gevaar vormen voor onze samenleving.

Hun doel is duidelijk: door het systematisch inzetten van desinformatie wil men tweedracht zaaien, het vertrouwen in onze democratische instellingen ondergraven en zo onze hele samenleving destabiliseren en verzwakken.» aldus het verslag.

Nog een stap verder zijn de *deepfakes* die klaarblijkelijk in «*real time*» kunnen toegepast worden, zoals een filter. Zo was er het voorval met burgemeesters van verschillende Europese hoofdsteden die misleid waren tot het voeren van videogesprekken met een *deepfake* van hun ambtgenoot in Kiev, Vitali Klitschko in juni 2022.

De burgemeester van Berlijn, Franziska Giffey, nam deel aan een gepland gesprek op het videoconferentieplatform *Webex* met een persoon die volgens haar leek en klonk als Klitschko (11).

«Er waren geen aanwijzingen dat het videoconferentiegesprek niet met een echte persoon werd gevoerd», zei ze in een verklaring.

Pas na ongeveer vijftien minuten, toen de vermeende burgemeester van Kiev aan de andere kant van de lijn begon te praten over het probleem van Oekraïense vluchtelingen die misbruik maken van Duitse uitkeringen en leek op te roepen om vluchtelingen terug te brengen naar Oekraïne voor militaire dienst, werd Giffey achterdochtig.

Toen de verbinding kort werd onderbroken, nam het kantoor van de Berlijnse burgemeester contact op met de Oekraïense ambassadeur in Duitsland, die via de autoriteiten in Kiev bevestigde dat de persoon in het videogesprek niet de echte Klitschko was, meldde de krant *Der Spiegel*.

Los van het feit dat *deepfakes* gebruikt worden om schade aan te richten is het nog schadelijker voor de democratie en het politieke bestel dat men alles in vraag

(11) *https://www.theguardian.com/world/2022/jun/25/european-leaders-deepfake-video-calls-mayor-of-kyiv-vitali-klitschko.*

au système politique. Si l'on ne peut même plus avoir confiance dans des images animées, en quoi peut-on encore avoir confiance?

Les *deepfakes* et le trucage de vidéos sont de plus en plus souvent employés pour ce que les professeurs de droit Danielle Citron et Bobby Chesney appellent le «dividende du menteur», c'est-à-dire la capacité qu'ont les personnes influentes de contester de manière crédible la véracité d'images embarrassantes. Ces personnes crient alors au *deepfake* ou à la manipulation pour jeter le discrédit sur une vidéo diffusée à leur insu et les présentant dans une situation compromettante ou pour s'en prendre à l'une des rares sources de protestation citoyenne dans les régimes autoritaires, en mettant en doute la crédibilité des images de violences d'État filmées à l'aide d'un *smartphone*. Ces agissements vont dans le prolongement des tromperies orchestrées par l'État dans les régimes autoritaires. Ainsi, au Myanmar, l'armée et les autorités ont, à plusieurs reprises, diffusé elles-mêmes des fausses images et mis en doute l'intégrité de preuves véritables de violations des droits humains (12).

Dans nos contrées, le danger serait, par exemple, que certains acteurs divulguent des images manipulées d'un président ou dirigeant mondial déclarant en apparence la guerre à un autre pays du monde, avec toutes les conséquences qui s'ensuivraient, ou que certains dirigeants soient contraints de démissionner à la suite de *deepfakes* les présentant, par exemple, comme impliqués dans l'un ou l'autre scandale.

Des *deepfakes* sont déjà utilisés comme armes dans la guerre qui oppose l'Ukraine et la Russie, comme cela a été souligné dans une publication récente du Centre d'excellence européen pour la lutte contre les menaces hybrides («*European Centre of Excellence for Countering Hybrid Threats*», basé à Helsinki), dont la Belgique est devenue le trentième partenaire en septembre 2021.

Le constat suivant a ainsi été dressé:

«La guerre de 2022 en Ukraine montre que l'utilisation de *deepfakes* est une tendance émergente et, même si les vidéos diffusées semblaient relativement peu raffinées, il ne faut pas oublier que nous sommes seulement au seuil d'une ère d'influence dominée par l'intelligence artificielle. L'OTAN, l'UE et les pays européens doivent par conséquent envisager de revoir en profondeur la doctrine et le processus d'influence et de contre-influence utilisant le domaine de l'information. L'utilisation de *deepfakes* en Ukraine a créé la nécessité de développer

(12) *https://www.wired.com/story/opinion-authoritarian-regimes-could-exploit-cries-of-deepfake/.*

begint te stellen. Als men zelfs bewegende beelden al niet meer kan vertrouwen, wat dan nog wel?

Deepfakes en videomanipulatie worden steeds vaker gebruikt voor wat de rechtenprofessoren Danielle Citron en Bobby Chesney het «leugenaarsdividend» noemen, het vermogen van invloedrijke personen om aannemelijke ontkenning te claimen voor belastend beeldmateriaal. Uitspraken als «Het is een *deepfake*» of «Het is gemanipuleerd» worden dan gebruikt om een gelekte video van een voor hen compromitterende situatie in diskrediet te brengen of om een van de weinige bronnen van burgerprotesten in autoritaire regimes aan te vallen, zoals de geloofwaardigheid van smartphonebeelden van staatsgeweld. Dit bouwt voort op de geschiedenis van door de Staat gesteunde misleiding in autoritaire regimes. Zo hebben in Myanmar het leger en de autoriteiten herhaaldelijk zelf nepbeelden gedeeld en de waarheidsgetrouwheid en integriteit van echte bewijzen van mensenrechtenschendingen in twijfel getrokken (12).

Het gevaar in onze gebieden bestaat erin dat bijvoorbeeld bepaalde actoren in de toekomst gemanipuleerde beelden zouden kunnen lekken van een president of een wereldleider die schijnbaar de oorlog aan een ander land verklaart, met alle gevolgen van dien. Of dat bepaalde leiders via dergelijke *deepfakes* tot ontslag gedwongen worden, omdat ze bijvoorbeeld zo in een of ander schandaal betrokken worden.

In de oorlog tussen Oekraïne en Rusland worden *deepfakes* reeds als wapen gebruikt. Dit blijkt uit een recente paper van het *European Centre of Excellence for Countering Hybrid Threats* (Helsinki), waar België in september 2021 de 30ste partner werd.

De volgende vaststelling werd er gedaan:

«De oorlog van 2022 in Oekraïne toont aan dat het gebruik van *deepfakes* een opkomende trend is, en ook al leken de uitgezonden video's vrij weinig geraffineerd, we mogen niet vergeten dat we pas aan de vooravond staan van het door AI-gestuurde beïnvloedingstijdperk. De NAVO, de EU en de Europese landen moeten daarom overwegen de doctrine en het proces voor beïnvloeding en tegenbeïnvloeding via het informatiedomein ingrijpend te actualiseren. Met het gebruik van *deepfakes* in Oekraïne ontstond de noodzaak om interne vermogens

(12) *https://www.wired.com/story/opinion-authoritarian-regimes-could-exploit-cries-of-deepfake/.*

des compétences internes pour détecter et contrer les *deepfakes* s'appuyant sur les technologies GAN. Dans un avenir proche, l'utilisation de technologies basées sur l'intelligence artificielle permettant de créer des séquences sonores et des vidéos manipulées risque dès lors de devenir une nouvelle norme dans les manières conventionnelles et non conventionnelles de faire la guerre. La formation d'un personnel spécialisé est indispensable pour acquérir les connaissances nécessaires à la lutte contre la guerre de l'information 3.0. (13)» (traduction)

C. Préjudice financier

Outre les aspects susmentionnés, les *deepfakes*, particulièrement lorsqu'ils sont combinés à des «voix de synthèse», c'est-à-dire à de faux enregistrements audios, peuvent constituer des outils dangereux aux mains d'imposteurs, de fraudeurs et d'arnaqueurs.

La fraude à l'identité, le chantage et les diverses formes de fraude sont plus que jamais facilités par la technologie du *deepfake*. Manipuler de manière ciblée le cours de certaines actions, voire porter préjudice à certaines marques ne relèvent plus de la fiction (14). Les *deepfakes* permettent de se faire passer pour quelqu'un d'autre de manière de plus en plus convaincante. Cela peut avoir de graves conséquences tant au niveau personnel qu'au niveau régional, national et même international.

te ontwikkelen om *deepfakes* op te sporen en tegen te gaan met behulp van GAN-technologieën. In de nabije toekomst zou het gebruik van op AI gebaseerde technologieën om gemanipuleerde spraak en video's te creëren dus een nieuwe norm kunnen worden in conventionele en niet-conventionele oorlogsvoering. De opleiding van toegewijd personeel is essentieel om de nodige kennis te verwerven over hoe informatieoorlog 3.0 kan worden tegengegaan (13).»

C. Financiële schade

Los van de bovenstaande aspecten kunnen *deepfakes*, in het bijzonder in combinatie met «synthetische stemmen», ofwel audio-*deepfakes*, een gevaarlijk hulpmiddel worden voor bedriegers, fraudeurs en oplichters.

Identiteitsfraude, afpersing en allerhande soorten fraude kan men door deze technologie beter dan ooit uitvoeren. Gericht de koersen manipuleren tot zelfs schade toebrengen aan bepaalde merken zijn niet meer ondenkbaar (14). Door *deepfakes* kan men zichzelf voor iemand anders uitgeven op een steeds overtuigendere manier. Zowel op persoonlijk, regionaal, nationaal tot zelfs op internationaal vlak kan dit grote gevolgen hebben.

Overview of different categories of risks associated with deepfakes

Psychological harm	Financial harm	Societal harm
• (S)extortion • Defamation • Intimidation • Bullying • Undermining trust	• Extortion • Identity theft • Fraud (e.g. insurance/payment) • Stock-price manipulation • Brand damage • Reputational damage	• News media manipulation • Damage to economic stability • Damage to the justice system • Damage to the scientific system • Erosion of trust • Damage to democracy • Manipulation of elections • Damage to international relations • Damage to national security

Source: *https://www.europarl.europa.eu/RegData/etudes/STUD/2021/690039/EPRS_STU(2021)690039_EN.pdf.*

Bron: *https://www.europarl.europa.eu/RegData/etudes/STUD/2021/690039/EPRS_STU(2021)690039_EN.pdf.*

(13) *https://www.hybridcoe.fi/wp-content/uploads/2022/06/20220623-Hybrid-CoE-Paper-14-AI-based-technologies-WEB.pdf*
(14) *https://www.europarl.europa.eu/RegData/etudes/STUD/2021/690039/EPRS_STU(2021)690039_EN.pdf.*

(13) *https://www.hybridcoe.fi/wp-content/uploads/2022/06/20220623-Hybrid-CoE-Paper-14-AI-based-technologies-WEB.p*df.
(14) *https://www.europarl.europa.eu/RegData/etudes/STUD/2021/690039/EPRS_STU(2021)690039_EN.pdf.*

D. Côtés positifs des *deepfakes*

Les *deepfakes* n'ont pas que des côtés négatifs. La technologie est neutre en soi, c'est la manière dont elle est utilisée qui lui donne son sens.

Toutes les personnes qui ont déjà utilisé un smartphone moderne pour prendre des photos auront sans doute pu expérimenter les avantages des technologies *deepfake* élémentaires. Les applications photo sont souvent pourvues de filtres de beauté, qui adaptent automatiquement les images. Des *deepfakes* plus évolués qui échangent des visages entiers ou modifient les voix peuvent également être réalisés légitimement, par exemple pour formuler un commentaire critique, faire une satire ou une parodie ou tout simplement divertir un public. D'autres possibilités d'application utile de la technologie *deepfake* peuvent être d'emblée imaginées dans les domaines des productions audio-graphiques, des interactions homme-machine, des vidéoconférences, de la satire, de l'expression créative personnelle ou encore dans le cadre des recherches et traitements médicaux.

L'innovation technologique qui va de pair avec cette technologie pousse également à une meilleure compréhension de ce domaine par et au sein de nos services de sécurité, ainsi qu'à l'élaboration de solutions par nos entreprises et à la mise à l'épreuve de cette technologie dans des jardins d'essais/«bacs à sable».

Dans le domaine de l'art et du divertissement grand public, les *deepfakes* connaissent également un succès grandissant. Lors de la finale de l'émission de divertissement «*America's got talent*», la technologie deepfake développée par la société limbourgeoise Metaphysic a ainsi permis de rendre vie à l'artiste légendaire qu'était Elvis Presley. Pendant le spectacle, le chanteur mort a été ressuscité, on lui a fait chanter deux chansons. De même, les membres du jury ont soudainement commencé à faire les choristes alors qu'ils étaient en réalité restés assis à leur table (15).

III. *DEEPFAKES* ET APPLICATION

A. Formation, éducation aux médias et vérificateurs de faits

Chaque mesure que l'on prend pour lutter contre l'utilisation néfaste des *deepfakes* arrive toujours trop tard. Les *deepfakes* sont en général perçus négativement car ils relèvent d'une technologie nouvelle encore méconnue du grand public et liée à des concepts comme la

(15) *https://www.vrt.be/vrtnws/nl/2022/09/14/mag-je-zomaar-iemand-tot-leven-brengen-met-een-deepfake/.*

D. Positieve kanten *deepfakes*

Niet alles is negatief rond *deepfakes*. De technologie is op zich neutraal, de manier waarop men het aanwendt is hoe men het betekenis geeft.

Iedereen die wel eens een moderne smartphone heeft gebruikt voor fotografie, heeft waarschijnlijk wel eens de voordelen ondervonden van elementaire *deepfake*-technologieën. Vaak zijn camera-apps uitgerust met schoonheidsfilters, die beelden automatisch aanpassen. Meer geavanceerde *deepfakes* waarbij hele gezichten worden verwisseld of spraak wordt aangepast, kunnen ook rechtmatig worden gemaakt om bijvoorbeeld kritische commentaar, satire en parodieën te leveren of gewoon om een publiek te vermaken. Andere voor de hand liggende mogelijkheden voor nuttig gebruik van *deepfakes* zijn in de context van audio-grafische producties, mens-machine-interacties, videoconferenties, satire, persoonlijke creatieve expressie en medische behandeling of onderzoek.

De technologische innovatie die gepaard gaat met deze technologie noopt ook tot het faciliteren van een beter begrip hiervan door en bij onze veiligheidsdiensten, én voor het uitwerken van oplossingen door onze ondernemingen en het testen van deze technologie via proeftuinen/sandboxes.

Op het gebied van kunst en mainstream entertainment zijn *deepfakes* ook aan een opmars bezig. Zo werd in de finale van het entertainmentprogramma «America's got talent», de legendarische artiest Elvis Presley terug tot leven gewekt mede door deepfake technologie. Dankzij het Limburgse bedrijf Metaphysic kon dit bewerkstelligd worden. Tijdens de act wekte men de dode zanger opnieuw tot leven en liet hem twee liedjes zingen. Ook de juryleden begonnen plots mee te doen als achtergrondzangers, terwijl ze in werkelijkheid gewoon achter de jurytafel zaten (15).

III. *DEEPFAKES* EN HANDHAVING

A. Scholing, mediawijsheid en *factcheckers*

Elke maatregel die men neemt tegen schadelijk gebruik van *deepfakes* komt nooit te vroeg en is altijd te laat. Over het algemeen hebben *deepfakes* een negatieve reputatie omdat het nog een nieuwe technologie is, nog niet al te bekend is onder de bredere bevolking en

(15) *https://www.vrt.be/vrtnws/nl/2022/09/14/mag-je-zomaar-iemand-tot-leven-brengen-met-een-deepfake/.*

manipulation, la désinformation, les infox et la vengeance pornographique.

Or les *deepfakes* présentent aussi des côtés positifs, notamment quand ils sont utilisés dans le secteur du divertissement ou dans le secteur artistique. Il n'en reste pas moins qu'il faut prémunir les citoyens d'aujourd'hui et de demain contre les risques potentiels que ces images manipulées recèlent.

À cet égard, il vaut mieux prévenir que guérir. La priorité est d'informer les citoyens que ce genre de choses existent, mais sans semer la peur pour autant. La première chose à faire est d'attirer l'attention sur le fait que la technologie peut être à la fois productive et destructrice.

À cet égard, il existe la méthodologie SIFT, qui promeut la vérification des sources (16). Elle est efficace non seulement contre les *deepfakes* mais aussi contre la désinformation générale et les infox.

veelal gelinkt wordt aan begrippen zoals manipulatie, desinformatie, *fake news* en wraakporno.

Het kent echter ook positieve kanten, zoals in de entertainmentsector of de kunstsector. Los hiervan is het belangrijk dat men de burgers van vandaag en morgen wapent tegen de mogelijke risico's die dergelijke gemanipuleerde beelden met zich meedragen.

In eerste instantie is voorkomen beter dan genezen. Het is nuttig burgers op de hoogte te brengen dat dergelijke dingen bestaan, zonder evenwel een angstcampagne te voeren. Het is echter in eerste instantie belangrijk om mee te geven dat technologie zowel productief als destructief kan zijn.

Zo is er bijvoorbeeld de SIFT-methodologie, die ertoe moet aanzetten om de bronnen te checken (16). Dit werkt niet enkel tegen *deepfakes* maar ook tegen algemene desinformatie en *fake news*. De Nederlandstalige benaming voor de SIFT- methodologie is het HALT-model.

La crise de la Covid-19 a confirmé que la fracture numérique (l'accès au matériel informatique et aux appareils numériques, la maîtrise des logiciels et les aptitudes des utilisateurs) était importante. Ce fossé s'est creusé encore davantage avec le passage accéléré aux outils numériques dans toutes les franges de la population.

Le développement de l'éducation aux médias et des compétences numériques est donc une nécessité non seulement pour les jeunes, mais aussi pour la société dans son ensemble.

En dispensant des cours d'éducation aux médias à nos enfants et nos jeunes dans le cadre de l'école, nous les armons contre la manipulation et la désinformation.

De Covid-19-crisis heeft bevestigd dat de digitale kloof (de toegankelijkheid tot informaticamateriaal en digitale apparatuur, beheersing van software en gebruikersvaardigheden) groot is. Die kloof werd nog groter door een versnelde overschakeling op digitale instrumenten in alle lagen van de bevolking.

De ontwikkeling van mediawijsheid en digitale vaardigheden is daarom niet alleen een noodzaak voor jongeren, maar een noodzaak voor de gehele samenleving.

Via lessen mediawijsheid in het onderwijs wapenen we onze kinderen en jongeren tegen manipulatie en desinformatie. Andere instellingen in permanente vorming

(16) *https://hapgood.us/2019/06/19/sift-the-four-moves/.*

(16) *https://hapgood.us/2019/06/19/sift-the-four-moves/.*

D'autres établissements de formation continue sont actifs aussi dans le domaine de l'éducation aux médias (17). Il existe ainsi des initiatives comme «*Link in de kabel*» et «Ouvrir mon quotidien».

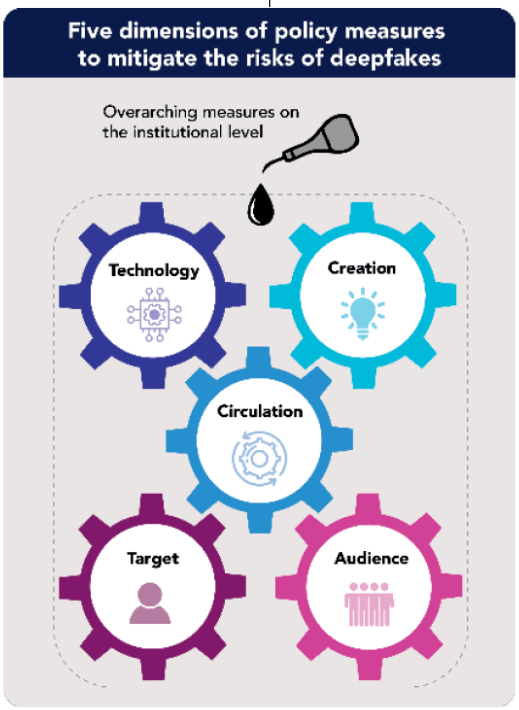
Les grands réseaux tels que *Facebook* et *Twitter* s'efforcent de filtrer eux-mêmes la désinformation et les *deepfakes* mais la contribution de vérificateurs de faits indépendants est aussi indispensable.

B. Politique et *deepfakes*

zijn ook actief op het gebied van mediavorming (17). Zo zijn er initiatieven zoals «Link in de kabel» en «*Ouvrir mon quotidien*».

Naast de grote kanalen, zoals *Facebook* en *Twitter* die desinformatie en *deepfakes* zelf proberen te filteren, is de input van onafhankelijke *factcheckers* onontbeerlijk.

B. Beleid en *deepfakes*



Source: *https://www.europarl.europa.eu/RegData/etudes/STUD/2021/690039/EPRS_STU(2021)690039_EN.pdf.*

Dans son étude, le Parlement européen identifie cinq domaines différents dans lesquels on doit agir si l'on veut s'attaquer au problème de l'incidence des *deepfakes* néfastes.

1) La dimension technologique

La dimension technologique comprend des options politiques orientées vers la technologie qui est à la base des *deepfakes* – des techniques d'apprentissage automatique qui s'appuient sur l'intelligence artificielle – et vers les acteurs qui sont impliqués dans la production et la fourniture de cette technologie. La régulation de cette technologie relève largement du cadre réglementaire de l'intelligence artificielle tel que proposé par la Commission européenne.

(17) *https://www.senate.be/informatieverslagen/7-110/Senaat_verslag_fake_news-2021.pdf.*

Bron: *https://www.europarl.europa.eu/RegData/etudes/STUD/2021/690039/EPRS_STU(2021)690039_EN.pdf.*

Het Europees Parlement duidt in zijn studie vijf verschillende domeinen aan waaraan gewerkt moet worden, wil men de impact van schadelijke *deepfakes* aanpakken.

1) Technologiedimensie

De technologiedimensie omvat beleidsopties die gericht zijn op de technologie die de basis vormt van *deepfakes* – op AI gebaseerde technieken voor machinaal leren - en de actoren die betrokken zijn bij de productie en levering van deze technologie. De regulering van dergelijke technologie valt grotendeels binnen het domein van het regelgevingskader voor AI zoals voorgesteld door de Europese Commissie.

(17) *https://www.senate.be/informatieverslagen/7-110/Senaat_verslag_fake_news-2021.pdf.*

Le cadre part d'une approche de la régulation de l'IA, fondée sur le risque. Dans la proposition de la Commission, les *deepfakes* sont désignés explicitement comme des «systèmes d'IA utilisés pour générer ou manipuler des images ou des contenus audio ou vidéo» devant répondre à certaines exigences minimales, notamment en matière de labellisation.

Ils ne relèvent pas de la catégorie «risque élevé», et il n'est pas certain qu'ils puissent relever de la catégorie «interdit». La proposition-cadre actuelle pour l'IA laisse donc une marge à l'interprétation. Étant donné que cette étude a analysé un large éventail d'applications de la technologie *deepfake*, dont certaines comportent clairement des risques élevés, il serait souhaitable de clarifier et de compléter la proposition-cadre relative à l'IA. Ainsi, on pourrait préciser quelles pratiques d'IA doivent être interdites sur la base du cadre réglementaire de l'IA, instaurer des obligations légales pour les fournisseurs de technologies *deepfake* et réglementer ces technologies en tant que technologies à haut risque (18).

Au Sénat, on avait déjà fait le lien antérieurement entre l'apprentissage automatique et les *deepfakes*. Ainsi, la dynamique de propagation permet de voir comment une rumeur se propage et ce, indépendamment de l'examen du contenu. Le lieu et la vitesse de l'échange d'informations diffèrent en effet selon qu'il s'agit de désinformation ou de contenus neutres.

L'entreprise *Textgain* – une *spin-off* de l'Université d'Anvers (UA) – utilise l'intelligence artificielle pour suivre de près ce qui se passe sur les médias sociaux. On peut ainsi déceler plus rapidement les propos haineux, la désinformation et la discrimination en ligne. L'entreprise essaie aussi de déterminer la manière dont les chambres d'écho se forment sur les médias sociaux.

On utilise aussi l'apprentissage automatique pour détecter les *deepfakes*. On n'en est pas encore au stade où l'IA supprime aussi des contenus; pour cela, une intervention humaine est toujours nécessaire (19).

2) La dimension de la création

Cette dimension comprend les options politiques qui visent à s'en prendre aux créateurs de *deepfakes* ou, dans la terminologie du cadre de l'IA, aux «utilisateurs» de systèmes d'IA. La proposition-cadre pour l'IA énonce déjà quelques règles et restrictions concernant

(18) *https://www.europarl.europa.eu/RegData/etudes/STUD/2021/690039/EPRS_STU(2021)690039_EN.pdf.*
(19) *https://www.senate.be/informatieverslagen/7-110/Senaat_verslag_fake_news-2021.pdf.*

Het kader gaat uit van een risicogebaseerde benadering van de regulering van AI. *Deepfakes* worden in het voorstel van de Commissie expliciet behandeld als «AI-systemen die worden gebruikt om beeld-, audio- of video-inhoud te genereren of te manipuleren», die aan bepaalde minimumeisen moeten voldoen, met name wat labeling betreft.

Zij vallen niet onder de categorie «hoog risico», en het blijft onzeker of zij onder de categorie «verboden» kunnen vallen. Het huidige kadervoorstel voor AI laat dus ruimte voor interpretatie. Aangezien dit onderzoek een breed scala aan toepassingen van *deepfake*-technologie heeft gedocumenteerd, waarvan sommige duidelijk een hoog risico inhouden, worden verduidelijkingen van en aanvullingen op het AI-kadervoorstel aanbevolen. Opties zijn onder meer een verduidelijking van welke AI-praktijken op grond van het AI-kader verboden moeten worden; het creëren van wettelijke verplichtingen voor aanbieders van *deepfake*-technologie en regulering van *deepfake*-technologie als zeer riskant (18).

In de Senaat werd eerder machinaal leren al in verband gebracht met *deepfakes*. Zo laat de propagatiedynamiek toe om na te gaan hoe een gerucht wordt verspreid, zonder de inhoud te bekijken. De locatie en snelheid van informatie-uitwisseling zijn immers anders bij desinformatie dan bij neutrale inhoud.

Het bedrijf *Textgain* – een *spin-off* van de Universiteit Antwerpen (UA) – houdt met behulp van artificiële intelligentie de vinger aan de pols van wat er leeft op sociale media. Hiermee kan men sneller haatspraak, desinformatie en discriminatie online opsporen. Het bedrijf probeert ook te achterhalen hoe echokamers zich vormen op sociale media.

Machinaal leren wordt eveneens gebruikt om *deepfakes* te detecteren. Het gaat nog niet zo ver dat AI ook inhoud verwijdert, daarvoor blijft er steeds een menselijke tussenkomst nodig (19).

2) Creatiedimensie

Deze dimensie omvat de beleidsopties die gericht zijn op het aanpakken van de makers van *deepfakes*, of in AI-kaderterminologie: de «gebruikers» van AI-systemen. Het voorstel voor het AI-kader formuleert reeds enkele regels en beperkingen voor het gebruik van

(18) *https://www.europarl.europa.eu/RegData/etudes/STUD/2021/690039/EPRS_STU(2021)690039_EN.pdf.*
(19) *https://www.senate.be/informatieverslagen/7-110/Senaat_verslag_fake_news-2021.pdf.*

l'utilisation de la technologie *deepfake*, mais des mesures complémentaires sont possibles. Parmi les options, il y a, notamment, la clarification des lignes directrices relatives à la méthode de labellisation, la limitation des exceptions à l'exigence de labellisation des *deepfakes* et l'interdiction générale de certaines applications.

Cette dimension concerne aussi ceux qui utilisent la technologie *deepfake* à des fins malveillantes: les «auteurs». Les utilisateurs malintentionnés de la technologie *deepfake* se retranchent souvent derrière l'anonymat et ne sont pas faciles à identifier, si bien qu'ils échappent à toute obligation en termes de responsabilité. On ne peut pas attendre de ces utilisateurs qu'ils se conforment volontairement à l'obligation de labellisation telle que prévue dans la proposition de cadre réglementaire pour l'IA.

3) La dimension de la circulation

Ce domaine comprend les options politiques qui visent à lutter contre la circulation de *deepfakes* en formulant des règles et des limitations pouvant être envisagées en ce qui concerne la diffusion de (certains) *deepfakes*. Les plateformes en ligne, les médias et les services de communication jouent un rôle crucial dans la diffusion de *deepfakes*.

La diffusion et la circulation d'un *deepfake* déterminent dans une large mesure l'ampleur et la gravité de son incidence. C'est pourquoi il est souvent recommandé de définir les responsabilités et les obligations des plateformes et des autres intermédiaires.

4) La dimension du groupe-cible

Les *deepfakes* malveillants ont des conséquences au niveau individuel pour les personnes mises en scène. Il ressort de l'étude européenne que les droits des victimes peuvent en principe être protégés, mais que cette protection est difficile à mettre en œuvre dans la pratique. Les auteurs de l'étude proposent dès lors différentes pistes pour améliorer la protection des victimes, notamment l'institutionnalisation de l'aide aux victimes de *deepfakes*, le renforcement de la capacité des autorités de protection des données à réagir face à l'utilisation de données à caractère personnel dans le cadre de *deepfakes* et la mise au point d'une approche uniforme en vue d'un bon usage des droits de la personnalité au sein de l'Union européenne.

5) La dimension publique

L'impact d'un *deepfake* dépasse le niveau individuel et peut avoir des répercussions jusqu'au niveau d'un

deepfake-technologie, maar aanvullende maatregelen zijn mogelijk. Opties zijn onder meer de verduidelijking van de richtsnoeren voor de wijze van labelen; beperking van de uitzonderingen op de vereiste *deepfake*-labeling en een algeheel verbod op bepaalde toepassingen.

Deze dimensie heeft ook betrekking op degenen die *deepfake*-technologie voor kwaadaardige doeleinden gebruiken: de «dader». Kwaadwillende gebruikers van *deepfake*-technologie verschuilen zich vaak achter de anonimiteit en kunnen niet gemakkelijk worden geïdentificeerd, waardoor zij aan de verantwoordingsplicht ontsnappen. Van deze gebruikers kan niet worden verwacht dat zij vrijwillig voldoen aan de labelingsverplichting die in het voorstel voor het AI-kader is opgenomen.

3) Circulatiedimensie

Dit domein omvat de beleidsopties die gericht zijn op het aanpakken van de circulatie van *deepfakes*, door het formuleren van mogelijke regels en beperkingen voor de verspreiding van (bepaalde) *deepfakes*. Onlineplatformen, media en communicatiediensten spelen een cruciale rol bij de verspreiding van *deepfakes*.

De verspreiding en circulatie van een *deepfake* bepaalt in grote mate de omvang en de ernst van de impact ervan. Daarom worden vaak verantwoordelijkheden en verplichtingen voor platformen en andere tussenpersonen aanbevolen.

4) Doelgroep-dimensie

Kwaadwillige *deepfakes* hebben gevolgen op individueel niveau, voor de personen die in de *deepfake* worden afgebeeld. Het Europees onderzoek heeft aangetoond dat de rechten van slachtoffers in beginsel kunnen worden beschermd, maar dat het vaak moeilijk blijkt om dit te bewerkstelligen. Daarom bieden ze verschillende opties aan om de bescherming van de slachtoffers te verbeteren, waaronder het institutionaliseren van de steun aan slachtoffers van *deepfakes*; het versterken van de capaciteit van gegevensbeschermingsautoriteiten om te reageren op het gebruik van persoonsgegevens voor *deepfakes* en het ontwikkelen van een uniforme aanpak voor het juiste gebruik van persoonlijkheidsrechten binnen de Europese Unie.

5) Publieksdimensie

De impact van een *deepfake* overstijgt het individuele niveau en kan doorwerken tot op groeps- of zelfs

groupe voire de la société. Cela dépendra en partie de la réaction du public: les gens croiront-ils le *deepfake*, relaieront-ils eux-mêmes le *deepfake* qu'ils reçoivent, perdront-ils confiance dans les institutions? La dimension publique constitue dès lors le dernier paramètre crucial que les décideurs politiques doivent prendre en compte pour atténuer les risques et les effets des *deepfakes*. Parmi les pistes citées à cet égard figurent la labellisation des sources fiables et l'investissement dans l'éducation aux médias et dans une citoyenneté responsable sur le plan technologique.

C. La technologie comme filtre contre les *deepfakes*

De nombreuses nouvelles études ont introduit diverses méthodes de détection de vidéos manipulées (*deepfake video-detection* – DVD). Certaines de ces méthodes revendiquent un taux de précision de plus de 99 % dans des cas spécifiques, mais de tels rapports de précision doivent être interprétés avec la prudence de rigueur. Le degré de difficulté de détection des vidéos manipulées varie sensiblement en fonction de divers facteurs, par exemple le taux de compression, la résolution d'image et la composition de l'ensemble de test (20).

Une analyse récente comparant les performances de sept détecteurs avancés sur cinq ensembles de données publics fréquemment utilisés rapportait un taux de précision très variable, allant de 30 à 97 %, sachant qu'aucun détecteur n'était significativement meilleur qu'un autre. La précision des détecteurs était la plupart du temps très variable pour les cinq ensembles de données tests. Les détecteurs ont généralement été conçus pour détecter un type précis de manipulation. Lorsqu'ils sont paramétrés sur de nouvelles données, ils affichent souvent une piètre performance. Il est donc vrai que beaucoup d'efforts sont déployés sur ce plan, mais on ne peut pas affirmer que certains détecteurs soient sensiblement plus performants que d'autres.

Quel que soit le taux de précision des détecteurs actuels, la détection de vidéos manipulées s'apparente à un jeu du chat et de la souris, où les progrès dans les méthodes de création de *deepfakes* tentent de rattraper les progrès réalisés dans les méthodes de détection. Pour qu'elles puissent offrir une protection efficace contre une diversité de *deepfakes* malveillants, les méthodes de détection doivent faire l'objet de multiples améliorations tentant d'anticiper la génération suivante de contenus manipulés.

(20) *https://insights.sei.cmu.edu/blog/how-easy-is-it-to-make-and-detect-a-deepfake/.*

maatschappelijk niveau. Of dit gebeurt, hangt gedeeltelijk af van de reactie van het publiek: zullen zij de *deepfake* geloven, *deepfakes* verder verspreiden wanneer zij die ontvangen, het vertrouwen in de instellingen verliezen? De publieksdimensie is daarom de laatste cruciale dimensie voor beleidsmakers om de risico's en effecten van *deepfakes* te beperken. Tot de hier opgesomde opties behoren het labelen van betrouwbare bronnen en het investeren in mediageletterdheid en verantwoord technologisch burgerschap.

C. Technologie als filter tegen *deepfakes*

Een grote hoeveelheid aan nieuw onderzoek heeft verschillende *deepfake* videodetectie (DVD) methoden geïntroduceerd. Sommige van deze methoden beweren een detectienauwkeurigheid van meer dan 99 % te hebben in speciale gevallen, maar dergelijke nauwkeurigheidsrapporten moeten met de nodige omzichtigheid worden geïnterpreteerd. De moeilijkheidsgraad bij het detecteren van videomanipulatie varieert sterk afhankelijk van verschillende factoren, waaronder de mate van compressie, de beeldresolutie en de samenstelling van de testset (20).

Een recente vergelijkende analyse van de prestaties van zeven geavanceerde detectoren op vijf openbare datasets die vaak worden gebruikt, liet een breed scala aan nauwkeurigheden zien, van 30 tot 97 %, waarbij geen enkele detector significant beter was dan een andere. De detectoren hadden doorgaans een zeer uiteenlopende nauwkeurigheid voor de vijf testdatasets. Gewoonlijk zijn de detectoren afgestemd op een bepaald soort manipulatie. Wanneer deze detectoren op nieuwe gegevens worden ingesteld, presteren zij vaak niet goed. Het is dus waar dat er op dit gebied veel inspanningen worden geleverd, maar het is niet zo dat bepaalde detectoren veel beter zijn dan andere.

Ongeacht de nauwkeurigheid van de huidige detectoren, is DVD een kat-en-muisspel. Vooruitgang in detectiemethoden wordt afgewisseld met vooruitgang in deepfake-productiemethoden. Voor een succesvolle verdediging tegen een veelheid van kwaadwillende *deepfakes* moeten de DVD-methoden herhaaldelijk worden verbeterd door te anticiperen op de volgende generatie van *deepfake*-inhoud.

(20) *https://insights.sei.cmu.edu/blog/how-easy-is-it-to-make-and-detect-a-deepfake/.*

Les méthodes d'hypertrucage connaîtront probablement une évolution rapide des possibilités de production de vidéos toujours plus dynamiques. La plupart des méthodes actuelles produisent des vidéos statiques en ce sens qu'elles mettent en scène des sujets immobiles avec un éclairage constant et un arrière-fond fixe. Mais les *deepfakes* de demain afficheront un dynamisme dans l'éclairage, la posture et l'arrière-fond. Les caractéristiques dynamiques de ces vidéos risquent de mettre à mal les modèles de détection actuels. Il est tout aussi préoccupant de savoir que les *deepfakes* dynamiques pourraient paraître plus crédibles à l'œil humain. Ainsi, une vidéo d'un dirigeant étranger roulant dans une voiturette de golf donnerait l'impression de refléter plus fidèlement la réalité qu'une vidéo du même dirigeant s'exprimant face à la caméra dans un studio statique.

IV. *DEEPFAKES* ET MESURES PRISES PAR D'AUTRES PAYS

A. Mesures prises en Belgique

Depuis décembre 2021, la Belgique dispose d'une Stratégie nationale de sécurité, qui servira désormais de cadre pour le développement d'une véritable culture de la sécurité mettant l'accent sur la création d'une capacité de résilience.

Cette stratégie accorde également une grande attention à la désinformation et à la lutte contre la désinformation:

«Dans le contexte du déploiement hybride de différents types de menaces, notre pays doit développer une résilience spécifique dans un certain nombre de domaines prioritaires tels que la désinformation (dans le cadre d'opérations d'information hostiles), les organisations violentes et extrémistes dans le domaine cybernétique, et la protection de notre potentiel scientifique et économique (la base de notre prospérité). Dans ce contexte, la stratégie de l'UE pour l'union de la sécurité fournit des orientations pour une approche intégrée et cohérente pleinement soutenue par la Belgique.

Cette stratégie vise, entre autres, à renforcer une coopération policière et judiciaire mieux intégrée au niveau européen, qui tient également compte de l'évolution du contexte technologique. Au niveau national, la mobilisation permanente contre les menaces hybrides nécessite des initiatives de détection, d'analyse, de synthèse et de coordination. Il est donc important que les structures nationales qui en sont chargées continuent à développer leurs capacités dans les domaines de la prévention, de la préparation, de la détection et continuent à développer une riposte.

Deepfake-methoden zullen waarschijnlijk snel uitbreiden qua mogelijkheden om video's te produceren die steeds dynamischer zijn. De meeste bestaande *deepfake*-methoden produceren video's die statisch zijn in de zin dat ze stilstaande onderwerpen weergeven met constante belichting en een onbeweeglijke achtergrond. Maar de *deepfakes* van de toekomst zullen dynamiek bevatten in belichting, houding en achtergrond. De dynamische kenmerken van deze video's kunnen de prestaties van bestaande *deepfake*-detectiemodellen aantasten. Even zorgwekkend is dat dynamische *deepfakes* geloofwaardiger zouden kunnen overkomen voor menselijke ogen. Zo zou een video van een buitenlandse leider die op een golfkarretje voorbijrijdt, er waarheidsgetrouwer uitzien dan wanneer diezelfde leider rechtstreeks tegen de camera zou spreken in een statische studio.

IV. *DEEPFAKES* EN MAATREGELEN IN ANDERE LANDEN

A. Maatregelen in België

Sinds december 2021 beschikt België over een Nationale Veiligheidsstrategie, die voortaan als kader moet dienen voor het ontwikkelen van een echte veiligheidscultuur, met nadruk op het creëren van weerbaarheid.

Hierbij wordt ook ruim aandacht geschonken aan desinformatie en de bestrijding hiervan:

«Ons land moet in de context van de hybride inzet van verschillende types dreigingen, specifieke weerbaarheid uitbouwen in een aantal prioritaire domeinen zoals desinformatie (als onderdeel van vijandelijke informatie operaties), cyber, gewelddadige en extremistische organisaties en de bescherming van ons wetenschappelijk en economisch potentieel (de basis van onze welvaart). In deze context biedt de EU-strategie voor de Veiligheidsunie een leidraad voor een geïntegreerde en coherente aanpak die België voluit steunt.

Deze strategie beoogt onder meer beter geïntegreerde politie en gerechtelijke samenwerking op Europees niveau die ook rekening houdt met de evolutie van de technologische context. Op nationaal niveau vereist de permanente mobilisatie tegen hybride dreigingen de nodige initiatieven voor detectie, analyse, synthese en coördinatie. Het is daarom van belang dat de nationale structuren die hiermee belast zijn hun capaciteit op het vlak van preventie, paraatheid, opsporing en reactie verder ontwikkelen.

Les menaces hybrides se caractérisant par l'utilisation combinée de méthodes et de techniques, une plateforme nationale et interdépartementale sur les menaces hybrides a déjà été mise en place afin de garantir une approche coordonnée reposant sur une large assise. Spécifiquement pour la désinformation, des travaux sont menés en vue de la mise en place d'un mécanisme interdépartemental pour la détection, la surveillance, l'analyse et le rapportage d'opérations de désinformation et d'information (21).»

Les mesures nationales précitées de lutte contre la désinformation encadrent donc aussi, en partie, l'approche en matière de contenus malveillants tels que les *deepfakes*.

B. Mesures prises en Europe

Actuellement, il n'y a pas de lois européennes ni de lois nationales au Royaume-Uni, en France ou en Allemagne qui visent à lutter spécifiquement contre les *deepfakes* (22). Toutefois, l'Union européenne est en train de rédiger des directives pour contrer l'abus des *deepfakes* au sein de l'UE.

Cela étant, on pourra dans un proche avenir, grâce à une version actualisée du *Digital Services Act* (DSA) de 2018, imposer aux grandes plateformes de médias telles que *Google* et *Facebook* l'obligation de lutter avec plus de fermeté contre de tels contenus malveillants. Le code de conduite volontaire, qui a été introduit en 2018, deviendra un système de corégulation, la responsabilité étant partagée entre les contrôleurs et les signataires du code de conduite.

Le code de conduite actualisé comprend des exemples de comportements de manipulation, comme les *deepfakes* et les faux comptes, contre lesquels les signataires devront lutter (23).

En outre, au niveau européen, la proposition de réglementation sur l'intelligence artificielle (*Artificial Intelligence Act)* devrait bientôt aboutir.

Avec cette proposition, la présidente de la Commission, Ursula von der Leyen, respecte l'engagement politique qu'elle avait annoncé dans le cadre de ses orientations

(21) *https://www.premier.be/sites/default/files/articles/NVS_Online_NL.pdf.*

(22) *https://f.datasrvr.com/fr1/320/16758/1207330_-_GMCQ_-_Spring_2020_Deepfakes.pdf.*

(23) *https://www.reuters.com/technology/google-facebook-twitter-will-have-tackle-deepfakes-or-risk-eu-fines-sources-2022-06-13/.*

Gezien hybride dreigingen gekenmerkt worden door het gecombineerd gebruik van methodes en technieken, werd reeds een nationaal en interdepartementaal platform «hybride dreigingen» opgericht, om een breed gedragen en gecoördineerde aanpak te garanderen. Specifiek voor desinformatie is werk lopende tot het opzetten van een interdepartementaal mechanisme voor het opsporen, monitoren, analyseren en rapporteren van desinformatie- en informatie-operaties (21).»

De bovengenoemde nationale maatregelen tegen desinformatie omkaderen dus ook voor een gedeelte de aanpak omtrent misleidende materialen zoals *deepfakes*.

B. Maatregelen in Europa

Er zijn momenteel geen Europese wetten of nationale wetten in het Verenigd Koninkrijk (VK), Frankrijk of Duitsland die specifiek gericht zijn op de aanpak van *deepfakes* (22). De Europese Unie is echter wel bezig met richtlijnen uit te werken inzake de aanpak van misbruik van *deepfakes* binnen de EU.

Via een bijgewerkte versie van de DSA (*Digital Services Act*) van 2018 kan men in de nabije toekomst grote mediaplatformen zoals *Google* en *Facebook* wel dwingen om strenger op te treden tegen dergelijk misleidend materiaal. De vrijwillige gedragscode, die in 2018 werd ingevoerd, wordt een coreguleringsregeling, waarbij de verantwoordelijkheid wordt gedeeld tussen de toezichthouders en de ondertekenaars van de gedragscode.

De bijgewerkte gedragscode bevat voorbeelden van manipulatief gedrag, zoals *deepfakes* en nepaccounts, die de ondertekenaars zullen moeten aanpakken (23).

Daarnaast zou op EU-vlak binnenkort de *Artificial Intelligence Act* moeten landen.

Met dit voorstel komt Commissievoorzitter Ursula Von der Leyen de politieke verbintenis na die zij in haar politieke richtsnoeren voor de Commissie 2019-2024

(21) *https://www.premier.be/sites/default/files/articles/NVS_Online_NL.pdf.*

(22) *https://f.datasrvr.com/fr1/320/16758/1207330_-_GMCQ_-_Spring_2020_Deepfakes.pdf.*

(23) *https://www.reuters.com/technology/google-facebook-twitter-will-have-tackle-deepfakes-or-risk-eu-fines-sources-2022-06-13/.*

politiques pour la Commission 2019-2024 «*Une Union plus ambitieuse*», à savoir la présentation, par la Commission européenne, d'une proposition législative en vue d'une approche européenne coordonnée relative aux implications humaines et éthiques de l'intelligence artificielle.

Le titre IV de cette proposition porte sur certains systèmes d'IA dont il faut tenir compte en raison des risques spécifiques de manipulation qu'ils présentent.

C'est ainsi que des obligations de transparence s'appliqueront aux systèmes qui i) interagissent avec les humains, ii) sont utilisés pour détecter des émotions ou déterminer l'association avec des catégories (sociales) sur la base de données biométriques, ou iii) génèrent ou manipulent des contenus (trucages vidéo ultraréalistes). Lorsque des personnes interagissent avec un système d'IA ou que leurs émotions ou caractéristiques sont reconnues par des moyens automatisés, elles doivent en être informées. Si un système d'IA est utilisé pour générer ou manipuler des images ou des contenus audio ou vidéo afin de produire un résultat qui ressemble sensiblement à un contenu authentique, il devrait être obligatoire de déclarer que le contenu est généré par des moyens automatisés, sauf pour certaines finalités légitimes faisant l'objet d'exceptions (domaine répressif, liberté d'expression). Cette obligation laisse la possibilité aux personnes de prendre des décisions en connaissance de cause ou de se désengager d'une situation donnée (24).

En ce qui concerne le caractère «à haut risque» ou non: qu'en est-il du double usage? Qu'en est-il du lien entre la création et la détection de *deepfakes*? Dans l'annexe III de la proposition, les systèmes d'IA à haut risque sont définis comme étant «*tous les systèmes utilisés à des fins de «predictive policing» et de «predictive justice», les détecteurs de mensonges ou systèmes similaires destinés à détecter les émotions, les systèmes destinés à détecter les hypertrucages («deep fake»), les systèmes destinés à évaluer la fiabilité des preuves et les systèmes destinés à être utilisés pour le profilage dans le contexte de la police ou de la justice* (25)».

C. Mesures prises par les États-Unis d'Amérique

Au cours des cinq prochaines années, le ministère américain de la Sécurité intérieure (*Department of Homeland Security* – DHS) publiera un rapport annuel sur les *deepfakes*. Ce rapport abordera toutes les formes

(24) *https://eur-lex.europa.eu/legal-content/FR/TXT/HTML/?uri=CELEX:52021PC0206&from=EN.*
(25) *https://www.oranedecontrole.be/files/DA210029_Avis_F.pdf.*

«Een Unie die de lat hoger legt» heeft aangekondigd, namelijk dat de Europese Commissie wetgeving zal voorleggen voor een gecoördineerde Europese aanpak van de menselijke en ethische implicaties van AI.

Titel IV van dit voorstel heeft betrekking op bepaalde AI-systemen waarmee rekening moet worden gehouden gezien de specifieke risico's op manipulatie die zij inhouden.

Er zullen volgens dit voorstel transparantieverplichtingen moeten gelden voor dergelijke systemen die i) interageren met mensen, ii) worden gebruikt om emoties te detecteren of mensen in te delen in (sociale) categorieën op basis van biometrische gegevens, of iii) inhoud genereren of manipuleren («*deep fakes*»). Wanneer personen interageren met een AI-systeem of hun emoties of kenmerken door geautomatiseerde hulpmiddelen worden herkend, moeten zij daarvan op de hoogte worden gebracht. Als een AI-systeem wordt gebruikt om beeld-, audio- of videomateriaal te genereren of te manipuleren dat een merkbare gelijkenis vertoont met authentieke inhoud, moet het verplicht zijn om bekend te maken dat de inhoud door geautomatiseerde hulpmiddelen is gegenereerd, behoudens uitzonderingen voor legitieme doeleinden (rechtshandhaving, vrijheid van meningsuiting). Dit geeft personen de mogelijkheid weloverwogen keuzes te maken of afstand te nemen van een bepaalde situatie (24).

Met betrekking tot het al dan niet «*high risk*»: wat met *dual use*? Wat met de link tussen creatie en detectie van *deepfakes*? In de derde bijlage van het voorstel van AIA wordt als high risk beschouwd: «*alle systemen die te maken hebben met «predictive policing» en «predictive justice», leugendetectors of gelijkaardige systemen die emoties proberen te detecteren, systemen om deep fakes te detecteren, systemen om de betrouwbaarheid van bewijsmateriaal te beoordelen en systemen om aan politionele of justitiële profiling te doen*» (25).

C. Maatregelen in de Verenigde Staten (VSA)

Het Amerikaanse ministerie van Binnenlandse Veiligheid (*Department of Homeland Security's* – DHS) brengt de komende vijf jaar jaarlijks een verslag uit over *deepfakes*. In zo'n verslag moeten alle mogelijke

(24) *https://eur-lex.europa.eu/legal-content/NL/TXT/HTML/?uri=CELEX:52021PC0206&from=EN.*
(25) *https://www.oranedecontrole.be/files/DA210029_Avis_F.pdf.*

de préjudices possibles causés par la technologie, depuis les campagnes d'ingérence étrangère jusqu'à la fraude en passant par les dommages causés à des groupes de population spécifiques.

En outre, le DHS est tenu, en vertu de la loi, de mener des études sur les technologies permettant de créer des *deepfakes* et sur les solutions possibles pour les détecter et les limiter. Enfin, toujours selon la loi, le ministère américain de la Défense doit examiner l'éventualité que des opposants créent des contenus *deepfake* visant des militaires américains ou des membres de leur famille et recommander des changements de politique.

La loi dénommée «*Identifying Outputs of Generative Adversarial Networks Act*» a été signée par le président Trump fin décembre 2020. Elle prévoit que la *National Science Foundation* doit mener des recherches sur la technologie *deepfake* et les mesures d'authentification, que le *National Institute of Standards and Technology* doit soutenir le développement de normes relatives aux *deepfakes* et que ces deux agences doivent développer des méthodes pour collaborer avec le secteur privé sur les capacités d'identification des *deepfakes* (26).

V. QUE PROPOSONS-NOUS?

«Un mensonge aura déjà parcouru la moitié de la terre que la vérité mettra seulement ses chaussures», dit-on parfois. On le constate avec les *deepfakes*: avant que la vérité n'émerge, une information mensongère aura déjà amplement circulé. Il est beaucoup plus difficile de convaincre la population que des images mobiles ont été manipulées parce que cette notion n'est pas encore très ancrée dans les esprits. Par ailleurs, un autre risque est que les *deepfakes* sèment une confusion générale, ce qui finit par déclencher une apathie et nuit assurément au système démocratique.

Les *deepfakes* peuvent donc causer d'énormes dommages. Combinés à des infox, des *trolls* et de la désinformation d'État, ils forment un véritable cocktail toxique qui peut porter gravement atteinte à la démocratie. Prendre des mesures pour lutter contre de telles pratiques n'est donc pas un luxe superflu, c'est même une amère nécessité.

(26) *https://www.asisonline.org/security-management-magazine/latest-news/today-in-security/2021/january/U-S-Laws-Address-Deepfakes/.*

vormen van schade door de technologie aan bod komen, van buitenlandse beïnvloedingscampagnes tot fraude en schade bij specifieke bevolkingsgroepen.

Bovendien draagt de wet het DHS op onderzoek te doen naar technologie voor het creëren van «*deepfakes*» en naar mogelijke oplossingen voor detectie en beperking. Ten slotte moet het Amerikaanse ministerie van Defensie volgens de wet de mogelijkheid bestuderen dat tegenstanders *deepfake*-inhoud creëren waarin Amerikaans militair personeel of hun familieleden zijn afgebeeld, en beleidswijzigingen aanbevelen.

De *Identifying Outputs of Generative Adversarial Networks Act* werd eind december 2020 door president Trump ondertekend. Deze wet vereist dat de *National Science Foundation* onderzoek doet naar *deepfake*-technologie en authenticiteitsmaatregelen, vereist dat het *National Institute of Standards and Technology* de ontwikkeling van normen met betrekking tot *deepfakes* ondersteunt en draagt beide agentschappen op om manieren te ontwikkelen om met de particuliere sector samen te werken aan *deepfake*-identificatiecapaciteiten (26).

V. WAT STELLEN WIJ VOOR?

«Een leugen kan half de wereld rondreizen terwijl de waarheid nog zijn schoenen aan het aantrekken is», wordt soms gezegd. Met *deepfakes* is de leugen de wereld al twee keer rondgetrokken, terwijl de waarheid nog niet uit zijn bed is gekomen. Het is veel moeilijker om de bevolking ervan te overtuigen dat bewegende beelden gemanipuleerd zijn omdat deze notie op dit ogenblik nog niet ingeburgerd is. Daarnaast loopt men ook het risico dat *deepfakes* algehele verwarring zaaien, wat uiteindelijk apathie uitlokt en het democratische bestel zeker niet ten goede komt.

De schade die men kan toebrengen met *deepfakes* is dus immens. Als men dit nog verder combineert met *fake news*, trollen en staatsgerichte desinformatie heeft men een ware giftige cocktail in handen die het hart van de democratie diep kan treffen. Het is dan ook geen overbodige luxe om maatregelen te nemen tegen dergelijke praktijken, het is zelfs een bittere noodzaak.

(26) *https://www.asisonline.org/security-management-magazine/latest-news/today-in-security/2021/january/U-S-Laws-Address-Deepfakes/.*

Avec la présente proposition de résolution, les auteurs espèrent faire un premier pas dans la lutte contre les risques découlant de cette nouvelle technologie et la limitation de ceux-ci.

La proposition de résolution présente un caractère transversal. Les TIC (technologies de l'information et de la communication), la politique scientifique, la protection de la vie privée et la sécurité sont des matières transversales.

*
* *

De indieners van dit voorstel van resolutie hopen een eerste stap te zetten in het tegengaan en beperken van de risico's die uit deze nieuwe technologie voortvloeien.

Dit voorstel van resolutie heeft een transversaal karakter. ICT (*information and communication technology*), wetenschapsbeleid, privacy en veiligheid zijn transversale aangelegenheden.

*
* *

PROPOSITION DE RÉSOLUTION

Le Sénat,

A. considérant que les *deepfakes* sont le summum actuel en termes de manipulation de supports visuels, avec de lourdes conséquences tant sur le plan personnel que sur le plan sociétal ou géopolitique;

B. considérant qu'il est facile de réaliser soi-même des *deepfakes*, notamment grâce à une variété d'applications, et que cela requiert de moins en moins de connaissances techniques;

C. considérant que les *deepfakes* audio peuvent imiter des voix existantes et représentent un danger encore plus grand lorsqu'ils sont combinés à des *deepfakes* visuels;

D. considérant que les *deepfakes* présentent majoritairement un caractère pornographique et que les *deepnudes* et *deepfakes* pornographiques prennent les femmes pour cibles de manière disproportionnée;

E. considérant que les *deepnudes* vont souvent de pair avec la vengeance pornographique;

F. considérant que dans un contexte politique, les *deepfakes* peuvent causer des dommages additionnels parce que l'authenticité des images ne peut pas être vérifiée à première vue;

G. considérant que des phénomènes tels que la désinformation et les *deepfakes* occupent une place de plus en plus importante dans les médias et dans la politique tant au niveau national qu'au niveau international;

H. considérant que des régimes autoritaires recourent systématiquement à la désinformation dans le but de semer la discorde, de saper la confiance dans nos institutions démocratiques et, partant, de déstabiliser et de fragiliser notre société dans son ensemble;

I. considérant que la problématique des *deepfakes* et des *deepnudes* en particulier, ainsi que de la vengeance pornographique, a déjà été évoquée à la Chambre des représentants à plusieurs reprises il y a peu de temps;

J. considérant que les *deepfakes* et le trucage de vidéos sont de plus en plus souvent employés pour ce que l'on appelle le «dividende du menteur», à savoir la capacité qu'ont des personnes influentes de contester de manière crédible la véracité d'images embarrassantes;

VOORSTEL VAN RESOLUTIE

De Senaat,

A. overwegende dat *deepfakes* de huidige spreekwoordelijke overtreffende trap zijn op het gebied van getrukeerd beeldmateriaal, met grote gevolgen op zowel persoonlijk, maatschappelijk als geopolitiek vlak;

B. overwegende dat *deepfakes* mede door een verscheidenheid aan *apps* makkelijk zelf te maken zijn en steeds minder technische kennis vereisen;

C. overwegende dat audio-*deepfakes* bestaande stemmen kunnen imiteren en een nog groter gevaar vormen indien ze gecombineerd worden met visuele *deepfakes*;

D. overwegende dat het grootste gedeelte van *deepfakes* pornografisch is en *deepnudes* en pornografische *deepfakes* disproportioneel vrouwen viseren;

E. overwegende dat *deepnudes* en wraakporno vaak samenvallen;

F. overwegende dat *deepfakes* in een politieke context extra schade kunnen aanrichten omdat de authenticiteit van de beelden niet op het eerste gezicht kan worden geverifieerd;

G. overwegende dat verschijnselen als desinformatie en *deepfakes* op nationaal en internationaal niveau steeds nadrukkelijker aanwezig zijn in de media en in de politiek;

H. overwegende dat autoritaire regimes door het systematisch inzetten van desinformatie tweedracht willen zaaien, het vertrouwen in onze democratische instellingen ondergraven en zo onze hele samenleving destabiliseren en verzwakken;

I. overwegende dat *deepfakes* en *deepnudes* in het bijzonder recentelijk reeds verschillende keren in de Kamer van volksvertegenwoordigers werden aangehaald, samen met wraakporno;

J. overwegende dat *deepfakes* en videomanipulatie steeds vaker gebruikt worden voor wat het «leugenaarsdividend» wordt genoemd, namelijk het vermogen van invloedrijke personen om aannemelijke ontkenning te claimen voor belastend beeldmateriaal;

K. considérant que les *deepfakes*, particulièrement lorsqu'ils sont combinés à des «voix de synthèse» ou à de faux enregistrements audios, peuvent constituer des outils dangereux aux mains d'imposteurs, de fraudeurs et d'arnaqueurs;

L. considérant que la formation, l'éducation aux médias ainsi que les vérificateurs de faits sont indispensables pour lutter contre la désinformation et les *deepfakes*;

M. considérant que dans son étude, le Parlement européen identifie cinq domaines différents dans lesquels on doit agir si l'on veut s'attaquer à l'incidence des *deepfakes* néfastes;

N. considérant que, quelle que soit la précision des détecteurs actuels, la détection des vidéos manipulées (*deepfake video-detection* – DVD) s'apparente à un jeu du chat et de la souris;

O. considérant que l'Union européenne (UE) est en train d'élaborer des directives sur la manière de lutter contre l'utilisation abusive des *deepfakes* au sein de l'UE;

P. considérant que les États-Unis ont adopté une législation qui promeut la recherche sur les *deepfakes*, l'intelligence artificielle (IA), les réseaux antagonistes génératifs (*generative adversarial networks* – GAN), l'apprentissage automatique et d'autres technologies apparentées;

Q. considérant qu'il est beaucoup plus difficile de convaincre la population que des images mobiles ont été manipulées parce que cette notion n'est pas encore bien ancrée dans les esprits,

Demande à tous les gouvernements compétents en la matière:

1) de promouvoir le développement d'un esprit critique permettant autant que possible aux citoyens de faire la distinction entre des deepfakes et des images réelles;

2) d'être attentifs à la violence fondée sur le genre et au cyberharcèlement en ligne, étant donné que les *deepfakes* et les *deepnudes* visent les femmes de manière disproportionnée, et de prendre aussi les mesures nécessaires en la matière;

3) de développer des modèles de menace en matière de *deepfakes* et de coupler les solutions y afférentes à une perspective tant régionale et nationale qu'internationale;

K. overwegende dat *deepfakes*, in het bijzonder in combinatie met «synthetische stemmen», ofwel audio-*deepfakes* een gevaarlijk hulpmiddel kunnen worden voor bedriegers, fraudeurs en oplichters;

L. overwegende dat scholing, mediawijsheid en *fact-checkers* onontbeerlijk zijn in de strijd tegen desinformatie en *deepfakes*;

M. overwegende dat het Europees Parlement in zijn studie vijf verschillende domeinen aanduidt waaraan gewerkt moet worden, wil men de impact van schadelijke *deepfakes* aanpakken;

N. overwegende dat ongeacht de nauwkeurigheid van de huidige detectoren, *deepfake* video-detectie (DVD) een kat-en-muisspel is;

O. overwegende dat de Europese Unie (EU) bezig is met richtlijnen uit te werken inzake de aanpak van misbruik van *deepfakes* binnen de EU;

P. overwegende dat men in de Verenigde Staten wetgeving heeft aangenomen die onderzoek naar *deepfakes*, artificiële intelligentie (AI), *Generatieve Adversariale Netwerken* (GAN), machineleren en andere verwante technologieën bevordert;

Q. overwegende dat het veel moeilijker is om de bevolking ervan te overtuigen dat bewegende beelden gemanipuleerd zijn, omdat deze notie op dit ogenblik nog niet ingeburgerd is;

Vraagt aan alle hiertoe bevoegde regeringen om:

1) in te zetten op de ontwikkeling van een kritische ingesteldheid van de mensen waardoor ze zoveel als mogelijk in staat zijn deepfakes te onderscheiden van werkelijke beelden;

2) aandacht te hebben voor online gendergerelateerd geweld en cyberpesten, aangezien *deepfakes* en *deepnudes* disproportioneel vrouwen raken en daartoe ook de nodige maatregelen te nemen;

3) dreigingsmodellen omtrent *deepfakes* te ontwikkelen en daarbij horende oplossingen te koppelen aan zowel een regionaal, nationaal als internationaal perspectief;

4) de maintenir un contact permanent avec des acteurs tels que les médias, les experts, les vérificateurs de faits et les plateformes de médias sociaux, afin d'avoir une meilleure connaissance de la menace que les *deepfakes* représentent et de parvenir à des solutions qualitatives;

5) de mettre en place des mécanismes de coordination appropriés entre les autorités, les médias, les plateformes de médias sociaux, les experts et le secteur de l'enseignement concernant l'utilisation et la production de médias de synthèse, en ce compris les *deepfakes*, les *deepnudes*, etc.;

6) d'organiser des campagnes d'information et de sensibilisation sur les *deepfakes* destinées à de larges groupes de la population, en particulier dans le cadre de la désinformation (politique), de l'escroquerie et de la vengeance pornographique;

7) d'analyser de manière plus approfondie la responsabilité incombant aux plateformes de médias qui diffusent pareils médias de synthèse ainsi que la responsabilité des créateurs et des fournisseurs qui vendent des logiciels capables de produire des *deepfakes*;

8) d'examiner dans quelle mesure les gouvernements peuvent promouvoir des normes éthiques, en particulier en ce qui concerne l'utilisation de *deepfakes* et de médias de synthèse, aussi bien dans des campagnes politiques et des campagnes des pouvoirs publics que dans des campagnes de la société civile;

9) de développer un cadre juridique clairement défini, permettent de poursuivre les auteurs d'utilisation abusive de deepfakes et d'assurer le respect de ce cadre.

Le 27 septembre 2022.

4) voortdurend in contact te staan met actoren zoals de media, experts, *factcheckers* en socialemediaplatformen, om zo een beter inzicht te krijgen in de dreiging omtrent *deepfakes* en te komen tot kwaliteitsvolle oplossingen;

5) passende coördinatiemechanismen op te zetten tussen de overheden, de media, de socialemediaplatformen, experten en onderwijs omtrent het gebruik en de productie van synthetische media, waaronder ook *deepfakes*, *deepnudes*, enz., vallen;

6) informatie- en sensibiliseringscampagnes te houden omtrent *deepfakes*, gericht tot brede bevolkingsgroepen, in het bijzonder in het kader van (politieke) desinformatie, oplichting en wraakporno;

7) dieper te analyseren welke verantwoordelijkheid de mediaplatformen hebben die dergelijke synthetische media verspreiden, alsook welke verantwoordelijkheid makers en aanbieders hebben die software verkopen die *deepfakes* kunnen produceren;

8) na te gaan in hoeverre de regeringen het promoten van ethische normen, in het bijzonder bij het gebruik van *deepfakes* en synthetische media, kunnen bevorderen bij zowel politieke campagnes, overheidscampagnes alsook campagnes van het maatschappelijk middenveld;

9) een duidelijk omschreven juridisch kader te ontwikkelen waardoor het misbruik van deepfakes kan vervolgd worden en de handhaving ervan te verzekeren.

27 september 2022.

Tom ONGENA.
Hélène RYCKMANS.
Fatima AHALLOUCH.
Gaëtan VAN GOIDSENHOVEN.
Karin BROUWERS.
Ludwig VANDENHOVE.
Chris STEENWEGEN.
Philippe DODRIMONT.

Imprimerie centrale – Centrale drukkerij