**Bart van der Sloot,
Yvette Wagensveld
en Bert-Jaap Koops**

DEEPFAKES:

DE JURIDISCHE UITDAGINGEN VAN EEN SYNTHETISCHE SAMENLEVING

November 2021,
Tilburg Institute for Law, Technology, and Society

Tilburg University

Deepfakes: De juridische uitdagingen van een synthetische samenleving

Bart van der Sloot,
Yvette Wagensveld
en Bert-Jaap Koops

Samenvatting

Een deepfake is beeld, geluid of ander materiaal dat geheel of gedeeltelijk is gefabriceerd of bestaand beeld, geluid of ander materiaal dat is gemanipuleerd met behulp van geavanceerde technische hulpmiddelen en dat niet of nauwelijks van echt te onderscheiden is. Deepfakes maken gebruik van Machine Learning technologie en Artificial Intelligence. Naast het ontdekken van patronen kunnen middels deze netwerken ook eenvoudig beelden en geluiden worden geproduceerd, die lijken en gebaseerd zijn op bestaand materiaal. Meerdere technologieën kunnen hiervoor worden ingezet, maar de meest populaire is gebaseerd op wat bekend staat als Generative Adversarial Networks (GAN) en Variational AutoEncoders. GAN heeft de grenzen van de moderne resultaten verlegd en de kwaliteit en de resolutie van de geproduceerde beelden verbeterd en loopt voorop als het gaat om betrouwbaar, met lage kosten en tijdsinvesteringen diverse beelden en geluiden te genereren met modellen die rechtstreeks leren uit bestaande data. Met deze techniek kan door middel van het bekijken van bijvoorbeeld duizend foto's van Donald Trump een nieuwe foto van Trump worden geproduceerd die niet een exacte kopie is van een van die duizend foto's, waardoor het een geheel nieuwe foto lijkt te zijn. Die toepassing geldt tevens voor audio en video. Met deze techniek kan binnen enkele minuten een filmpje worden gegenereerd waarin een persoon dingen lijkt te zeggen of te doen die hij in werkelijkheid nooit heeft geuit of gedaan.

Deepfakes kunnen op tal van manieren worden ingezet voor positieve doeleinden, zoals humor en satire, voor het opsporen van criminelen en het infiltreren in criminele netwerken, voor entertainmentdoeleinden zoals in games en films, voor medische toepassingen, voor het 'passen' van kleding in de retailsector en het geven van rondleidingen in musea. Daarnaast zijn er ook veel negatieve toepassingen, zoals het genereren van (kinder)porno, fraude en misleiding, haat zaaien en aanzetten tot geweld, het verspreiden van misinformatie en het beïnvloeden van democratische verkiezingen. Naast concrete gevolgen van dergelijke toepassingen kunnen deepfakes ook belangrijke maatschappelijke gevolgen in het algemeen hebben. Daarbij valt te denken aan een afnemend vertrouwen in de media, de democratie en de rechtsspraak en belemmeringen voor het functioneren van deze instituten door de hoeveelheid nepmateriaal dat wordt gecreëerd en verspreid. De vrees is dat doordat binnen enkele jaren het merendeel van de onlinecontent gemanipuleerd zal zijn, het steeds lastiger te verifieren is wat echt is en wat nep voor journalisten, voor rechters en voor de burger zelf. Ook hebben de deepfake pornotoepassingen niet zelden een negatief effect op vrouwen en hun maatschappelijke positie en vrezen experts voor nadelige effecten voor opgroeiende meisjes van wie een deepfake circuleert op school. Meer dan 95% van de deepfakes zijn pornografisch van aard en vaak zonder toestemming van de betrokkene gegenereerd.

Tegen deze achtergrond is de probleemstelling van dit onderzoek: *'Dienen huidige en toekomstige*

onrechtmatige of strafwaardige uitingsvormen van deepfaketechnologie te leiden tot aanpassingen van de bestaande wetten en regels (met name de Uitvoeringswet AVG, het burgerlijk procesrecht en straf(proces)recht), of is bestaande wetgeving toereikend?' Deze vraag is beantwoord door middel van een literatuurstudie, een juridische analyse, het houden van interviews met experts en het analyseren van wetgeving in andere landen ten aanzien van deepfakes of de gevolgen daarvan. Dit onderzoek richt zich daarbij primair op horizontale relaties; dat wil zeggen dat dit onderzoek zich primair heeft gericht op burgers als gebruikers/actoren van deepfakeberichten en slechts incidenteel op het gebruik van deze techniek door actoren zoals (grote) bedrijven of (grotere) groepen. Ook is de aandacht primair uitgegaan naar deepfakeberichten gericht op individuen of kleine groepen mensen.

Belangrijkste bevindingen

Het Nederlands strafrecht is over het algemeen goed toegerust om deepfakes aan te pakken die dusdanig kwalijk zijn dat ze als strafwaardig kunnen worden beschouwd. Dat geldt zowel voor deepfakes die als nieuw middel worden ingezet om bestaande strafbare feiten te plegen, als voor deepfakes die qua inhoud strafwaardig lijken. Twee aanpassingen zijn evenwel mogelijk. Ten eerste valt momenteel niet onder een strafbepaling het geval waarin deepfake-seksvideo's niet worden verspreid, maar puur voor eigen gebruik worden gemaakt en bekeken. Het is een rechtspolitieke vraag of het voor eigen gebruik maken van zulke deepfakes van iemand zonder diens toestemming strafbaar zou moeten worden gesteld. Ten tweede is er een mogelijke lacune tussen artikel 231a Sr, dat identiteitsfraude strafbaar stelt waar biometrische gegevens worden misbruikt in situaties waarin die gegevens identificatie tot doel hebben, en artikel 231b Sr, dat identiteitsfraude met niet-biometrische gegevens strafbaar stelt. Misbruik van biometrische gegevens in gevallen waarin die gegevens niet identificatie tot doel hebben, is niet strafbaar, omdat artikel 231b Sr beperkt is tot niet-biometrische gegevens. Mocht de wetgever het wenselijk achten om kwalijke deepfakes – met name deepfakes die civielrechtelijk onrechtmatig zijn maar geen specifiek strafbaar feit opleveren – ook strafrechtelijk aan te kunnen pakken, dan valt te overwegen artikel 231b Sr aan te passen door het schrappen van de clausule 'niet zijnde biometrische persoonsgegevens' in artikel 231b Sr, of door deze clausule te vervangen door 'in andere gevallen dan bedoeld in artikel 231a'.

Deepfakes lopen tegen een aantal obstakels aan onder het gegevensverwerkingsregime van de Algemene Verordening Gegevensbescherming. Er moet een legitieme verwerkingsgrond zijn. Allereerst kan worden geopteerd voor toestemming van degene die in de deepfake wordt afgebeeld; dit zal doorgaans slechts een optie zijn als diegene een bekende is van de maker van de deepfake. Als het gaat om een deepfake waarop geen gevoelige zaken zijn te zien, zoals seksuele handelingen, dan kan het ook gaan om het geval waarin de belangen die worden gediend met de deepfake groter zijn dan de belangen van het datasubject om niet geportretteerd te worden. Dit zou het geval kunnen zijn bij een onschuldige satirische video van een politicus. Toch blijkt reeds enkel uit dit vereiste hoe nauw de legitieme toepassingsmogelijkheden voor deepfakes binnen de AVG zijn. Daarbij komt de plicht om de geportretteerde ervan op de hoogte te stellen dat hij in een deepfake figureert. De vraag is daarbij of het datakwaliteitsbeginsel niet zo moet worden gelezen dat deepfakes per

definitie verboden zijn. Hetzelfde geldt voor de vereisten van doel en doelbinding, waaruit volgt dat gegevens in principe alleen voor het doel mogen worden verwerkt waarvoor ze initieel zijn verzameld. Deepfakes geven per definitie een onjuiste voorstelling van zaken en gegevens zoals foto's en video's worden zelden verzameld met het vooropgezette doel om daar een deepfake van te maken. Dan zijn er ook nog de diverse rechten van het datasubject waar rekening mee moet worden gehouden, zoals het recht op rectificatie en het recht om vergeten te worden.

Binnen het Europees Verdrag voor de Rechten van de Mens moet voor deepfakes worden gekeken naar het samenspel van artikel 8 EVRM, waarin het recht op privacy is vervat, en artikel 10 EVRM, waarin het recht op vrijheid van meningsuiting is vervat. Het Europees Hof voor de Rechten van de Mens heeft geoordeeld dat onder het recht op privacy ook valt het recht op de bescherming van de eer en goede name en van de reputatie. Ook heeft het Hof geoordeeld dat de vrijheid van meningsuiting zeer ruim moet worden begrepen en ook omvat het recht om te schokken, te beledigen en te verwarren. Bij deepfakes met een mogelijk onrechtmatig karakter zullen dus vaak twee partijen een beroep kunnen doen op twee verschillende mensenrechten: de maker van de deepfake op zijn recht op vrijheid van meningsuiting, de afgebeeldene op zijn recht op eer en goede naam en recht op reputatie. Omdat het Hof weinig algemene regels stelt en iedere individuele zaak op zijn eigen merites, met het oog op de omstandigheden van het geval, beoordeelt, kan niet in algemene zin worden gezegd hoe deze twee rechten zich bij deepfaketoepassingen tot elkaar verhouden. Dit zal per zaak moeten worden bekeken.

Uit de landenstudie blijkt een diversiteit aan benaderingen van deepfakes en daaraan gerelateerde onderwerpen als des- en misinformatie. In China is het gebruik van deepfake- en virtual reality-technologieën voor de productie en verspreiding van desinformatie/misinformatie en nepnieuws, zowel door aanbieders van audio-videodiensten als door hun gebruikers, verboden. Ook moeten aanbieders audio-/video-informatie controleren en filteren als onrechtmatige content wordt gevonden. Zodra aanbieders van netwerkdiensten informatie of inhoud aantreffen die illegaal is, moeten zij de verwerking daarvan stopzetten en de verdere verspreiding blokkeren. In de Verenigde Staten hebben drie staten tot dusver wetgeving vastgesteld om het probleem van het zonder toestemming creëren en verspreiden van expliciet seksueel materiaal aan te pakken - Californië, Virginia en New York. Terwijl Virginia het zonder toestemming maken en verspreiden van seksueel expliciete deepfakes strafbaar heeft gesteld, heeft Californië wetgeving aangenomen die personen die daarin worden afgebeeld een privaatrechtelijke grond tot het instellen van een vordering biedt. Interessant is dat de wetgeving van New York ook voorziet in een rechtsvordering wegens ongeoorloofd commercieel gebruik van deepfakes, gemaakt met gebruikmaking van de beeltenis van een overleden uitvoerende kunstenaar. Een aantal staten heeft regels gesteld ten aanzien van het verspreiden van nepinformatie ten tijde van verkiezingen. Zo stelt een Texaanse wet het maken en publiceren van materiaal dat is bedoeld om de uitslag van een verkiezing te beïnvloeden, strafbaar.

Voor dit onderzoek zijn vijftien interviews afgenomen. Elf interviews zijn gehouden met internationale experts, vier met Nederlandse

experts op het gebied van het procesrecht. Hun verwachting is dat het gebruik van deepfaketechnologie de komende jaren een grote vlucht zal nemen. Zij voorspellen dat over zo'n zes jaar meer dan 90% van alle digitale content in meer of mindere mate is gemanipuleerd. Niet alleen is het volgens de geïnterviewden bijna onmogelijk om met het blote oog vast te stellen of een video of ander materiaal een deepfake is of niet, ook technische detectiemethoden hebben hun grenzen. De beste detectietechnieken die nu bestaan kunnen slechts zo'n 65% van de deepfakes ontdekken, de andere 35% glipt door het net. De verwachting van experts is dat de mogelijkheid om via technische middelen deepfakes te ontdekken eerder af dan toe zal nemen.

Bovendien wijzen zij erop dat ook het omgekeerde probleem zal ontstaan: het is vrij eenvoudig om met deepfaketechnologie op bestaand en niet gemanipuleerd materiaal sporen (artefacten) van manipulatie achter te laten, die de detectie-technologie kan ontdekken. De detectie-technologie zal het materiaal dan aanmerken als fake en mogelijk blokkeren, terwijl het om authentiek materiaal gaat. Bovendien is het probleem dat dergelijke technieken meestal 'waarheids-' of 'betrouwbaarheidspercentages' geven. Dan is bijvoorbeeld de uitkomst: de kans dat deze video authentiek, dat wil zeggen niet gemanipuleerd, is, is 78%. Als 90% van de online content op termijn geheel of gedeeltelijk gemanipuleerd is, detectiemethoden slechts een deel van de gemanipuleerde content kunnen ontdekken en zelfs dan slechts een waarschijnlijkheidspercentage kunnen geven dat content al dan niet gemanipuleerd is, dan roept dit volgens de geïnterviewden grote vragen op en problemen voor het functioneren van de rechtstaat, de democratie en de nieuwsvoorziening.

Algemene kader

Ten eerste kunnen deepfakes grote gevolgen hebben voor het vertrouwen in de media, het functioneren van de rechtsstaat en van de democratie; ook kunnen ze in algemene zin een negatieve impact hebben op de sociale en maatschappelijke positie van vrouwen. Naast deze grotere, meer maatschappelijke gevaren zijn er ook specifieke, kwalijke toepassingen van deepfaketechnologie. Een deepfake-pornofilm kan een catastrofale impact hebben op de professionele carrière van een vrouw, haar sociale positie en haar zelfbeeld; in extreme gevallen kan dit tot zelfmoord leiden. Deepfakes worden misbruikt voor het plegen van fraude en misleiding. Dit kan gaan om financieel gewin, ook kunnen deepfakes worden ingezet om bedrijfsgeheimen te ontfutselen of politieke besluitvorming te beïnvloeden of te frustreren. Daarnaast kunnen deepfakes worden ingezet om aan te zetten tot haat en geweld, bijvoorbeeld tegen minderheden en kunnen ze worden gebruikt om de intellectuele eigendomsrechten van artiesten te omzeilen en te ondermijnen.

Ten tweede zijn de mogelijke positieve toepassingen van deepfaketechnologie met name te vinden binnen professionele relaties, zoals tussen klant en bedrijf (bijvoorbeeld binnen de retailsector), patiënt en arts, burger en politicus, sekswerker en klant, werknemers van verschillende nationaliteit die met elkaar vergaderen en toepassingen binnen de entertainmentindustrie. Deze studie heeft maar één veelvoorkomende positieve toepassing van deepfaketechnologie in burger-burgerrelaties geïdentificeerd en dat is de inzet voor satire.

Ten derde is techniek nimmer neutraal. Bepaalde toepassingen worden gefaciliteerd of mogelijk gemaakt door het ontwerp van een technologie, anderen afgeremd of onmogelijk gemaakt. Dat is van belang omdat uit onderzoek blijkt dat meer dan 95% van de deepfakes wordt gebruikt voor zogenoemde *non-consensual porn*, het vervaardigen van pornografisch materiaal over iemand zonder diens toestemming. Daarbij moet worden opgeteld het gebruik van deepfaketechnologie voor fraude, misleiding en het verspreiden van schadelijk nepnieuws en het feit dat deepfakes, door de toenemende verwarring tussen fictie en werkelijkheid die zij per definitie veroorzaken, een negatieve impact kunnen hebben op het vertrouwen in de media, de rechtstaat en de democratie en het bestaan van een gedeelde werkelijkheid. De verwarring tussen feit en fictie is instrinsiek aan deze techologie en zal zich ook manifesteren bij positieve toepassingen. Zelfs die toepassingen hebben dus altijd een nadelig bijeffect.

Ten vierde is tegelijkertijd van belang dat deepfakes tot nu toe worden ingezet op een wijze die aansluit bij maatschappelijke tendensen die toch al zichtbaar zijn. De onderliggende problemen zijn breder en maatschappelijk van aard. Deepfake pornofilmpjes zijn in feite een uitvloeisel van het disrespect voor vrouwen en het objectiveren van het vrouwenlichaam dat zowel offline en zeker online hoogtij viert. Deepfake misinformatie past in het *post-truth* tijdperk, waarin meningen belangrijker worden dan feiten en waarin steeds meer groepen in hun eigen bubbel en waarheid leven. Het gebruik van deepfake voor politieke doeleinden sluit aan bij een toename aan interstatelijke vijandelijkheden via digitale wegen, die zich ook uiten in tal van hacks en spionageactiviteiten.

Tot slot is wellicht het belangrijkste inzicht dat regulering niet kan volstaan met aanpassingen in het materieel recht en het procesrecht op specifieke punten, hoewel sommige aanpassingen zeker mogelijk en wellicht wenselijk zijn. Het belangrijkste probleem ten aanzien van deepfakes in horizontale verhoudingen en meer in het algemeen van privacyschendingen in horizontale verhoudingen is gelegen in het toezicht op en de naleving van het vigerende recht. De meeste problematische toepassingen van deepfakes zijn al verboden of juridisch ingekaderd: het kernprobleem is daarom niet de wetgeving zelf, maar de handhaving daarvan.

Reguleringsopties

Deze studie heeft een breed palet aan mogelijke reguleringsopties gegeven. Op basis van rechtspolitieke keuzes kan de wetgever bepalen welke reguleringsopties wenselijk en haalbaar zijn. Sommige opties zijn direct invoerbaar, anderen zien op de lange termijn. De reguleringsopties moeten in onderlinge samenhang worden bezien. Soms adresseren meerdere opties eenzelfde onderliggend probleem. De keuze voor de ene optie betekent dan dat andere opties achterwege kunnen blijven.

Materieel recht		Procesrecht	Handhaving/toezicht
Reguleringsoptie 1: Voer een debat over de vraag of het maken of bezitten van "intrinsiek" strafwaardige (moreel verwerpelijke) deepfakes strafbaar zou moeten worden gesteld.	Reguleringsoptie 5: Bekijk in hoeverre de herziening van de huishoudelijke exceptie binnen het gegevens-beschermingsrecht wenselijk is	Reguleringsoptie 9: Bekijk in hoeverre er wet- en regelgeving of beleid dient te worden ontwikkeld of aang-escherpt om (deep)fake bewijsmateriaal in de rechtszaal tegen te gaan	Reguleringsoptie 10: Bekijk in hoeverre het wenselijk is om wet- of regelgeving te ontwik-kelen waarin het ver-vaardigen, aanbieden, dowloaden of gebruiken van deepfaketechnolo-gie of technologie die kan worden ingezet om deepfakes te genereren wordt verboden
Reguleringsoptie 2: Overweeg artikel 231b Sr aan te passen door de clausule 'niet zijnde biometrische gegevens' te schrappen of aan te passen.	Reguleringsoptie 6: Bekijk inhoeverre het aannemen van wet- of regelgeving waarin nadere regels worden gesteld met betrekking tot uitingen over publiek figuren wenselijk is		Reguleringsoptie 11: Bekijk in hoeverre het wenselijk is om wet- of regelgeving aan te nemen die ziet op de verplichte controle door internet-providers, burgers en/ of de AP van content op de authenticiteit en de rechtmatigheid voordat die content wordt gepub-liceerd en/of verspreid
Reguleringsoptie 3: Bekijk in hoeverre het creëren van wet- of regelgeving aangaande post-mortem privacy wenselijk is	Reguleringsoptie 7: Bekijk in hoeverre het wenselijk is om wet- of regelgeving te creëren ten aanzien van evident onware, onjuiste of mis-leidende uitingen		Reguleringsoptie 12: Start een publiekscam-pagne met informatie over de gevaren van deepfakes, waarin nieuwe sociale normen worden geëxpliciteerd en best practices worden benadrukt
Reguleringsoptie 4: Bekijk in hoeverre het creëren van wet- en regelgeving omtrent het gebruik van volledig door AI gegenereerde personen wenselijk is	Reguleringsoptie 8: Bekijk in hoeverre het wenselijk is om wet- of regelgeving te ontwik-kelen ten aanzien van het beïnvloeden van verkiezingen of politieke besluitvorming middels het creëren of verspre-iden van misinformatie		

Figuur 19: Overzicht van reguleringsopties