**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

Inhoud

klik op een titel om naar het desbetreffende hoofstuk te navigeren

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.

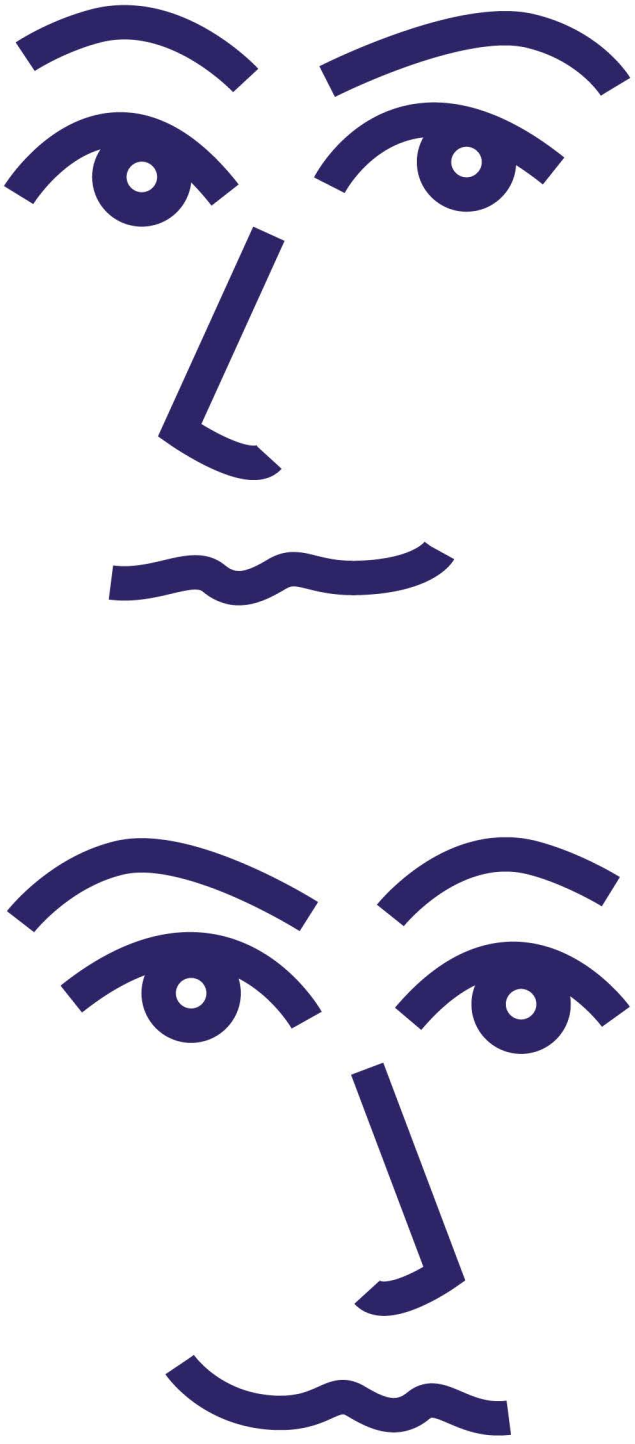
Tot slot is wellicht het belangrijkste inzicht dat regulering niet kan volstaan met aanpassingen in het materieel recht en het procesrecht op specifieke punten, hoewel sommige aanpassingen zeker mogelijk en wellicht wenselijk zijn. Het belangrijkste probleem ten aanzien van deepfakes in horizontale verhoudingen 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 aangescherpt om (deep)fake bewijsmateriaal in de rechtszaal tegen te gaan	Reguleringsoptie 10: Bekijk in hoeverre het wenselijk is om wet- of regelgeving te ontwikkelen waarin het vervaardigen, aanbieden, dowloaden of gebruiken van deepfaketechnologie 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 gepubliceerd 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 misleidende uitingen		Reguleringsoptie 12: Start een publiekscampagne 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 ontwikkelen ten aanzien van het beïnvloeden van verkiezingen of politieke besluitvorming middels het creëren of verspreiden van misinformatie		

Figuur 19: Overzicht van reguleringsopties



De schatting is dat over 5 jaar meer dan 90% van alle online content in een of andere vorm zal zijn gemanipuleerd. Bestaat er nog zoiets als een gedeelde werkelijkheid? En wat zijn de consequenties van een synthetische werkelijkheid voor het vertrouwen in elkaar, in de media en in de politiek?

1. Introductie

Dit rapport beschrijft de opkomst van deepfaketechnologieën, analyseert in hoeverre het huidige rechtssysteem toereikend is om de nadelige gevolgen daarvan te adresseren en geeft aanbevelingen om bestaande lacunes weg te nemen. Dit hoofdstuk geeft een korte introductie op de grotere belangen die met de opkomst van deepfakes op het spel staan (paragraaf 1.1), de aanleiding en afbakening van dit onderzoek (paragraaf 1.2), de probleemstelling en onderzoeksvragen (paragraaf 1.3), de gekozen methodologie (paragraaf 1.4) en de aanpak die voor dit onderzoek wordt gehanteerd, met daarbij een korte leeswijzer (paragraaf 1.5).

1.1 Deepfakes: maatschappelijke betekenis en belang

Hoe ziet een wereld eruit waarin fictie en werkelijkheid steeds moeilijker van elkaar te onderscheiden zijn, welke invloed heeft die vermenging op het functioneren van maatschappelijke instituties, op sociale cohesie en op onderlinge relaties? Door de opkomst van deepfakes zullen dit soort vragen op de politieke en maatschappelijke agenda komen. Deepfakes zijn gemanipuleerde beeld- en/of audiofragmenten die door middel van Artificial Intelligence (AI) in ettelijke minuten kunnen worden vervaardigd. De technologie is in een aantal jaar met rasse schreden in kwaliteit, snelheid en kostenefficiëntie vooruitgegaan, zozeer zelfs dat dergelijke nepvideo's, -foto's en -geluidsfragmenten al nauwelijks van echt te onderscheiden zijn. Deze trend zal zich naar verwachting doorzetten.

Dat heeft consequenties. Als over een jaar of tien een nog groter deel van de wereldbevolking in het bezit van is een smartphone, vrijwel iedereen toegang heeft tot het internet en selfies en filmpjes van alledaagse activiteiten worden gedeeld en verspreid via de diverse sociale netwerken, dan is het materiaal waarmee deepfakes kunnen worden vervaardigd in overdaad beschikbaar. De deepfaketechnologieën zullen nog beter en sneller zijn geworden en nog meer burgers zullen gratis of tegen een kleine betaling een deepfake app hebben geïnstalleerd. Zelfs in dit conservatieve scenario, waarin geen rekening wordt gehouden met revolutionaire technologieën of met hybride leefomgevingen die door middel van Virtual Reality kunnen worden gecreëerd, is duidelijk dat deepfakes een belangrijke maatschappelijke ontwikkeling teweegbrengt. Deepfakes zullen op beperkte en wellicht hinderlijke schaal worden ingezet, zoals een via de buurtapp verspreide deepfake waarin de impopulaire buurvrouw met een heksenhoed zetelt op een bezem[1], tot mensen die tijdelijk in een 'alternatieve' kersttoespraak van de Koningin trappen[2] tot ouders die ten onrechte geld overmaken omdat ze geloven dat hun kind in de problemen zit na het horen van een fakegeluidsfragment[3] en zelfs nog ter verificatie hebben video-gebeld met hun kind, maar in een programma zijn getrapt waarbij real-time nepbeelden en geluid kan worden geproduceerd.[4]

Deepfakes kunnen echter ook grotere maatschappelijke gevolgen hebben.[5] Daarbij valt bijvoorbeeld te denken aan het vertrouwen in de media en het ontstaan van parallelle werkelijkheden, trends die nu al waarneembaar zijn en versterkt kunnen worden door middel van deepfakes. Als er steeds meer deepfakes op het internet verschijnen dat wordt het steeds diffuser welke berichten echt en welke nep zijn en zullen bepaalde berichten, al dan

niet fake, binnen bepaalde groepen en online omgevingen worden gedeeld, terwijl andere daarbuiten worden gelaten. Het onderstreept des te meer de vraag wie de taak heeft om de 'waarachtigheid' of juistheid van berichten te controleren: de burger, traditionele media, de internetplatformen waarop de berichten worden gedeeld? Stel in 2031 hebben deepfakes zich zo ontwikkeld als nu de verwachting is, ze zijn kwalitatief hoogstaand en makkelijk toegankelijk voor burgers en iemand post een sensationeel filmpje op Twitter waarin te zien is hoe een aantal blanken mannen de Koran ontheiligen. Binnen no-time genereert het filmpje aandacht: Joop.nl spreekt er schande van, racistische uitlatingen van anonieme reaguurders volgen, cabaretiers maken er grappen over. Maar hoe beoordeel je als burger of het filmpje waarheidsgetrouw is? Je weet inmiddels hoe eenvoudig het is om zelf een deepfake te maken en hebt misschien zelf wel eens mensen bij de neus genomen, bovendien is het vaker voorgekomen dat een mediarel achteraf onterecht bleek, omdat het om een onecht bericht ging. Negeer je het nieuws, probeer je zelf een inschatting te maken van de waarachtigheid of juistheid, of vertrouw je alleen berichten die via traditionele media zijn verschenen en door hen geverifieerd?[6] Voor deze media legt dit een extra druk, omdat het vertrouwen in dit instituut reeds tanende is,[7] mensen nog meer dan nu de bronnen kunnen selecteren die hun eigen wereldbeeld bevestigen[8] en een zogenoemde 'reality-fatigue' kan optreden.[9] Als media meermaals per week content krijgen aangeboden die na analyse onjuist bleek te zijn, zullen zij zich wellicht genoodzaakt zien om zowel vanwege de kosten als de tijd die dergelijke analyse vergt geheel af te zien van het werken met materiaal dat niet van beroepsmatige journalisten afkomstig is, wat betekent dat ook een deel van de waardevolle content wordt genegeerd.[10] Bovendien is een mogelijk gevolg dat de discussie over geverifieerde content zich alsnog voltrekt, maar dan via kanalen die minder strikte regels hanteren. Het oude principe van *publish or perish* kan media dan voor een lastig dilemma[11] plaatsen. Moeten sociale media alle content op waarheidsgetrouwheid controleren[12] en zo ja, hoe dan en tegen welke prijs? Daarbij geldt het probleem dat als zij met standaardregels en algoritmes werken, er uit voorzorg ook publicaties zullen worden geblokkeerd die wel waardevol en waarachtig zijn.[13]

Ook in de rechtszaal kan de opkomst van deepfakes een grote impact hebben.[14] Het zal immers eenvoudig zijn om ofwel voor de grap ofwel om iemand doelbewust schade te berokkenen, handelingen te laten verrichten of woorden te laten spreken die strafbaar of onrechtmatig zijn. De nu al grote last op het Openbaar Ministerie om content op hun eventuele waarachtigheid en strafbaarheid te toetsen zal alleen maar toenemen. Deepfakes roepen tevens de vraag op waar de bewijslast dient komen te liggen en hoe hoog de lat moet zijn. Moet het Openbaar Ministerie aantonen dat een bepaald beeld- of geluidsfragment waarachtig is, of is het aan de verdediging om aan te tonen dat dit niet zo is? Als er alleen een percentage kan worden gegeven van de kans dat een bepaald beeld- of audiofragment waarachtig is, waar ligt dan de grens? Als het deels aan de burger is om aan te tonen dat bepaalde content waar of onwaar is, kan dat mogelijk leiden tot een ongelijkheid in de rechtspositie, aangezien de kosten voor een gedegen verificatie van video- en geluidsfragmenten significant kunnen zijn. En hoe zit het met de verdeling van bewijslast en de drempel voor toelaatbaar bewijs in civielrechtelijke procedures, bijvoorbeeld als een burger de eer en

goede naam van een andere burger lijkt te hebben geschonden of in een voogdijzaak waarin de moeder een geluidsfragment van een agressieve vader inbrengt, die de waarachtigheid daarvan betwist.[15] Dit kan niet alleen ten gevolg hebben dat rechtszaken langer duren en nog complexer worden wegens bewijslastkwesties, ook blijft er altijd een 'plausible deniability' over. Een veroordeelde kan altijd volhouden dat het bewijs dat tegen hem is gebruikt gefabriceerd was.

Daarnaast kunnen kwaadwillenden deepfakes inzetten, niet alleen om particulieren of bedrijven geld of bedrijfsgeheimen te ontfutselen,[16] ook kunnen ze worden ingezet om de financiële markten te manipuleren, bijvoorbeeld door nepberichten over omvallende banken in omloop te brengen, of een mogelijke concurrent uit te schakelen.[17] Deepfakes worden nu al ingezet om haat of geweld jegens minderheden aan te jagen,[18] om politieke rivalen zwart te maken[19] en om buitenlandse mogendheden te destabiliseren, bijvoorbeeld als door middel van nepberichtgeving het verkiezingsproces wordt beïnvloed[20] of als er video's verschijnen van de hoogste militair die ogenschijnlijk de oorlog verklaart aan een ander land.[21] Niet voor niets zijn het juist Inlichtingen- en Veiligheidsdiensten die waarschuwen voor de gevaren van deepfaketechnologie en desinformatie van buitenlandse actoren op Nederlands grondgebied.[22]

Als laatste voorbeeld van de meer maatschappelijke gevolgen van deepfakes kan er op worden gewezen dat veruit de meeste toepassingen van deepfakes momenteel seksueel van aard zijn en dat daarvan met name vrouwen het slachtoffer zijn. Daarbij gaat het zowel om publieke personen als om direct bekenden. Er is met name vrees voor wat dit voor gevolgen heeft voor jonge meiden op de middelbare school.[23] Nu al zijn de schattingen dat 20 tot wel 60% van de tienermeiden slachtoffer wordt van slutshaming[24] en andere ongewenste seksuele intimiteiten;[25] dergelijke praktijken via sociale media geuit wordt ook wel slutshaming 2.0 genoemd, omdat niet alleen de schaal en intensiteit exponentieel toenemen, maar het ook steeds moeilijker wordt om aan een bepaald beeld of imago te ontkomen.[26] Deepfakes zullen deze ontwikkeling alleen maar versterken en kunnen een slutshaming 3.0 met zich meebrengen,[27] bijvoorbeeld door jongens die met behulp van bestaande beeltenissen en geluidsfragmenten van hun ex een pornofilm zo manipuleren dat het pubermeisje in kwestie daarin lijkt te figureren.[28] Zelfs is bekend dat nu al niet alleen jongeren onderling dergelijke expliciete content van elkaar maken, maar dat zelfs ouders naaktbeelden van klasgenootjes fabriceren en verspreiden.[29] Experts vrezen niet alleen voor schade op individueel niveau, die zeer groot kan zijn,[30] maar dat dit op termijn kan betekenen dat jonge meiden zich altijd bewust zijn van een mogelijk gevaar en hun gedrag uit voorzorg zullen kuisen, wat fnuikend kan zijn voor hun persoonlijke ontwikkeling.[31]

1.2 Aanleiding en afbakening onderzoek

In de Kabinetsvisie Horizontale Privacy geeft de regering aan systematisch vooruit te willen kijken naar nieuwe technologieën die een risico met zich meebrengen voor de privacy van Nederlanders.[32] Dit onderzoek naar deepfakes past in die strategie, omdat alhoewel deepfaketechnologie momenteel nog in de kinderschoenen staat, de verwachting is dat deepfakes op termijn grote maatschappelijke consequenties kunnen hebben. Het anticiperen door middel van wetgeving en beleid op deze

gevolgen past ook in het bredere internationale kader waarin van de Verenigde Staten[33] tot China[34] en van Europol[35] tot in gremia binnen de Navo[36] en de Verenigde Naties[37] wordt opgeroepen tot regels, kaders en expliciete verboden voor bepaalde toepassingen van deze technologie.

Dit onderzoek zal een analyse bieden van het fenomeen deepfake, de mogelijke kansen en risico's van deze technologie in kaart brengen, beoordelen in hoeverre het huidige rechtssysteem afdoende waarborgen biedt, identificeren waar er juridische lacunes bestaan en aangeven waar aanpassingen aan vigerende wet- en regelgeving mogelijk zijn om deze lacunes te adresseren.

Daarbij geldt wel een belangrijke begrenzing, namelijk dat dit rapport zich met name zal richten op deepfakes gemaakt door burgers over andere burgers. De Kabinetsvisie waaruit dit onderzoek volgt kwam voort uit de startnotitie Horizontale Privacy van Kamerlid Koopmans en de discussie die daarop in de Tweede Kamer volgde. In zijn startnotitie stelt Koopmans dat er naast de zogenoemde verticale privacy – die ziet op de privacyschendingen van de overheid jegens de burger – en diagonale privacy – die ziet op privacyinbreuken begaan door ondernemingen en dan met name grote internet intermediairs tegen hun klanten en andere consumenten – ook aandacht moet zijn voor horizontale of onderlinge privacy – die ziet op privacyschendingen begaan door medeburgers, bijvoorbeeld met behulp van een drone of smartphone of door uitlatingen op sociale media.

'Online shaming, sextortion en internetpesten zijn aan de orde van de dag. Beelden zijn vaak voor eeuwig en kunnen wereldwijd worden gedeeld. Met gezichtsherkenningstechnologie, die nu ook voor privépersonen beschikbaar is, is straks niemand meer anoniem op straat. Een mobiele telefoon is ook een camera die wereldwijd kan uitzenden, ook zonder dat de gefilmde het weet. Verborgen camera's en afluisterapparatuur zijn niet alleen van de AIVD maar van iedereen. Met simpele locatie-trackers, microfoons en drones kan iedereen ongezien worden gevolgd, ieder gesprek worden opgenomen en ieder hoekje worden gefilmd. Door data-profilering en tracking-technologie is het privéleven van iedereen heel makkelijk in kaart te brengen, ook door medeburgers en kleine bedrijven. De situatie is fundamenteel anders dan vroeger. Er is een groot verschil tussen de buurman die tussen de schutting door gluurt en de buurman die tussen de schutting door *filmt*, want films zijn terug te zien, kunnen met iedereen worden gedeeld en gekoppeld aan andere persoonlijke informatie. Informatietechnologie heeft grote consequenties voor de manier waarop we leven. Naast de vele en belangrijke positieve gevolgen van efficiëntie, handel, verbondenheid en informatieverschaffing staan ook belangrijke negatieve aspecten zoals permanente observatie en individuele analyse. Dit kan gevolgen hebben voor hoe vrij mensen zich voelen, en soms ook voor hoe veilig ze daadwerkelijk zijn.'[38]

Uit deze aanleiding volgt dat in dit onderzoek met name stil zal worden gestaan bij de inzet van deepfakes in horizontale verhoudingen; dat wil zeggen dat dit onderzoek zich primair zal richten op burgers als gebruikers/actoren van deepfakeberichten en slechts incidenteel als het gaat om het gebruik van deze techniek door actoren zoals (grote) bedrijven of (grotere) groepen. Ook zal de aandacht primair uitgaan naar deepfakeberichten gericht op individuen of kleine groepen mensen. Uiteraard geldt

hierbij dat posities niet altijd helder zijn te onderscheiden. Een burger kan ook de eigenaar van een B.V. of eenmanszaak zijn of de leider van een groepering, vrijwel alle verspreidingen, ook in horizontale verhoudingen, vindt plaats via internet intermediairs als Facebook, Instagram en Tiktok en lang niet altijd is duidelijk of de maker van een deepfake een burger, een groep of een statelijke actor is. Bovendien zetten bijvoorbeeld statelijke actoren vaak burgers en groepen in, al dan niet in hun eigen of in een derde land, om destabiliserende content te verspreiden, omdat zij daar dan niet direct zelf op kunnen worden aangesproken. Alhoewel de grens tussen burger, groep, bedrijf en staat derhalve niet eenduidig is te trekken zal, zoals figuur 1 ook laat zien, er met concentrische cirkels van aandachtsgebieden worden gewerkt. In dat figuur worden indicatieve, hypothetische en niet-exhaustieve voorbeelden genoemd van toepassingen van deepfakes.

Actor \ Adressant	Burger	Groep	Organisatie	Staat
Burger	(Wraak)porno; afpersing; bedreiging; satire/memes van bekende Nederlanders of familieleden	Deepfakes door burgers die gehate groepen in diskrediet willen brengen, zoals etnische of religieuze minderheden; burgers die juist door middel van een deepfake Syrische immigranten welkom willen heten in hun taal	Burgers die organisaties in diskrediet willen brengen, bijvoorbeeld een Deepfake van Rupert Murdoch waarin hij stelt dat zijn kranten nooit over bosbranden mogen berichten	Burgers die (voor de grap of met serieuzere bedoelingen) politieke leiders (binnenlands of buitenlands) woorden in de mond leggen; burgers die het staatshoofd middels een ludieke deepfake feliciteren
Groep	Activisten die fakenews van Trump maken om steun te krijgen voor hun goede doel; jihadisten die door neponthoofdingen angst willen zaaien; kunstenaarscollectief dat middels een deepfake van een overleden schilder het volk wil verblijden	Deepfakes ter ondersteuning van een groepsbelang, bijvoorbeeld groep pro-pieten die door een fake video de anti-pieten in diskrediet wil brengen	Groep activisten die Shell of een andere organisatie in diskrediet wil brengen door een fakebericht, bijvoorbeeld van een groot olielek	Activisten uit Hongkong die deepfakes gebruiken om de Chinese staat/ politici in diskrediet te brengen; groep hackers die dreigt compromitterende deepfake van een kandidaat vlak voor de verkiezingsdag in omloop te brengen
Organisatie	Deepfakes om klanten te werven; fakenieuws om sentiment van de bevolking te manipuleren; inzetten deepfake voor het maken van video's, voor films of voor herinneringen aan overledenen; gebruik van deepfakes voor het passen van mode op realistisch zelfbeeld	Deepfake om groepen bang te maken, bijvoorbeeld deepfake van medisch bedrijf om steun te krijgen voor het opnemen van een bepaald medicijn in het basispakket door druk van specifieke patiëntengroep; deepfake om groepen patiënten te behandelen of te ondersteunen (bv alzheimer)	Deepfake om bedrijfsgeheimen/ middelen van een andere organisatie te ontfutselen; medewerkers van een Nederlandse en een Chinese organisatie die door middel van deepfaketechnologie live met elkaar kunnen spreken in elkaars taal	Deepfake om politieke besluitvorming te beïnvloeden, bijvoorbeeld het laten lijken alsof het hoofd van een oliebedrijf dat een boorvergunning wil politici omkoopt met steekpenningen
Staat	Door fakenieuws burgers (binnenlands of buitenlands) misleiden; in diskrediet brengen van dissidenten; het voorlichten van minderheidsgroepen door het staatshoofd in hun dialect of taal tot hen te spreken	Statelijke actoren die deepfakes gebruiken om minderheden in diskrediet te brengen (bijvoorbeeld Oeigoeren ogenschijnlijk radicaal religieuze boodschappen laten verkondigen)	Bijvoorbeeld Orban die een deepfake van Soros laat maken, of Erdogan van Gülen, om hun organisaties in diskrediet te brengen	Politieke leiders /hooggeplaatste militair van derde land onwaarachtige zaken laten beweren om land/democratie te destabiliseren; Rusland die deepfakes maakt om Oekraïne de schuld van de MH17-ramp in de schoenen te schuiven

Figuur 1: Categorisering deepfakes naar gebruiker en adressant: Groen is de kern van dit onderzoek, gevolgd door de toepassingen in blauw, dan die in geel en de vakken in rood zullen slechts zijdelings aan bod komen

Een tweede afbakening is dat dit onderzoek zich primair zal richten op de juridische en beleidscontext. Het rapport bevat geen uitputtende beschrijving van de technische achtergrond van deepfaketechnologieën, de werking daarvan en de te verwachte toekomstige toepassingen van Artificial Intelligence om deepfakes te detecteren. Ook bevat het geen economische analyse van de kosten van de technologie en de voordelen in termen van kostenreductie of het ontwikkelen van nieuwe bedrijfsmodellen. Technische, economische en sociale aspecten van deepfakes en de inzet daarvan komen in dit rapport aan bod in zoverre ze relevant zijn voor de juridische en beleidscontext. De aandacht zal primair uitgaan naar een bestudering van het bestaande juridische kader en de toepassing daarvan op deepfaketechnologieën. Is het bestaande kader afdoende om mogelijke nadelige gevolgen te voorkomen, biedt het voldoende ruimte om kansen te benutten en waar zijn mogelijk aanpassingen mogelijk en hoe zouden die er eventueel uit kunnen zien?

Een derde afbakening volgt uit deze juridische en beleidsmatige insteek. In zoverre er lacunes zullen worden gevonden in het bestaande juridische kader en daar oplossingsrichtingen voor zullen worden aangedragen zullen deze primair, maar niet uitsluitend, van juridische aard te zijn, terwijl niet elke juridische lacune per definitie ook juridisch geadresseerd hoeft te worden. Daarbij valt te denken te denken aan de ontwikkeling van technologieën die in staat zijn om deepfakes te herkennen, waarop statelijke actoren als Inlichtingen- en Veiligheidsdiensten en de grote internetondernemingen inzetten, en aan bewustwordingscampagnes, bijvoorbeeld op scholen of via de televisie, die de burger ervan moeten doordringen dat hij waakzaamheid moet betrachten en zich ervan bewust moet zijn dat een gedeelte van het nieuws en content die tot hem komen onecht zijn. Dit rapport zal primair bekijken in hoeverre wet- en regelgeving kan worden aangepast om de gevolgen van deepfakes op een adequate wijze te adresseren.

Een vierde en laatste afbakening betreft het fenomeen van deepfakes zelf. Er bestaat geen algemeen geaccepteerde definitie van deepfake. Zowel uit de bestudeerde literatuur als uit de voor deze studie gehouden interviews blijkt dat sommigen het concept zeer beperkt definiëren, bijvoorbeeld het vermengen van twee bestaande beelden door middel van *Generative Adversarial Networks,* terwijl anderen kiezen voor een bredere afbakening. Het gaat volgens hen om iedere vorm van manipulatie of fabricatie van audio, video of andere signalen door middel van Artificial Intelligence of Machine Learning. In dit rapport wordt gekozen voor een brede definitie, zowel omdat dit aansluit bij de wetenschappelijke en maatschappelijke tendens om bij (zeer) realistische maar geheel of gedeeltelijk gefabriceerde content te spreken over deepfakes als omdat de technische middelen waarmee beeldmanipulatie wordt uitgevoerd minder relevant is voor de uiteindelijk juridische en beleidscontext dan het resultaat. Daarvoor is veeleer van belang de mate waarin de gefabriceerde content reëel lijkt en/of voor waar wordt aangenomen.

Het beste is om een ideaaltype deepfake voor te stellen. De meest typische deepfake is een video die is gecreëerd met hoogwaardige technologische middelen waarin een bestaand persoon iets lijkt te doen of te zeggen dat diegene in werkelijkheid niet heeft gedaan of gezegd, waarbij het voor de consument van de video niet of nauwelijks mogelijk is deze manipulatie te ontwaren. Rond

dit kernvoorbeeld kunnen tal van toepassingen worden bedacht, die in meer of mindere mate perifeer zijn. Het kan gaan om video's die wel echt lijken, maar niet met hoogwaardige technologische middelen zijn gegenereerd; het kan gaan om hoogwaardige video's van niet bestaande personen die wel echt bestaand lijken; het kan gaan om video's die op een eerste oogopslag echt lijken, maar voor de oplettende kijker niet; het kan gaan om een fake-audiofragment of een gemanipuleerd satellietsignaal dat echt lijkt, maar minder 'direct binnenkomt' bij de consument dan een video; het kan gaan om kleine manipulaties, die soms inherent zijn aan de techniek, zoals het glad maken van de huid van een persoon of compressie op audioberichten; etc. Zes factoren zijn derhalve van belang:

- Type gegevensdrager en daarmee de mate waarin de content bij de gebruiker 'binnenkomt'
- De 'geavanceerdheid' van de gebruikte technologie
- De mate van manipulatie
- De mate waarin de manipulatie wezenlijk is voor die informatie die wordt overgebracht
- De vraag of de deepfake een bestaand of niet bestaand persoon betreft
- De mate waarin de gebruiker de content voor waar aanneemt

Vanuit deze benadering is de werkdefinitie die in dit rapport wordt gehanteerd van deepfakes:

> *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*

Voor de juridische context en beleidscontext zal ten aanzien van gemanipuleerde beelden mogelijk nog een verdere begrenzing nodig zijn. Niet iedere manipulatie zal noodzakelijkerwijs aan regulering dienen te worden onderworpen. Een filter, dat op steeds meer apparaten standaard is ingebouwd, waardoor de huid van personen wat egaler wordt gemaakt zal meestal niet problematisch zijn. Toch kan zelfs zo'n onschuldige manipulatie van belang zijn, bijvoorbeeld als een getuige een verdachte moet identificeren of als een dermatoloog middels een online consult een diagnose tracht te stellen. De vraag of een manipulatie, hoe klein ook, gereguleerd dient te worden is sterk contextafhankelijk. Het is niet altijd van tevoren, ten tijde van het vervaardigen van de deepfake, te voorspellen of een manipulatie in een later stadium kleine of grote gevolgen zal hebben.

Bij de in dit rapport gegeven voorbeelden zal een breed palet worden gegeven aan deepfakes. Daarbij zullen beelden die aanvankelijk deepfakes leken te zijn, maar dat later toch niet bleken te zijn of waaromtrent later onduidelijkheid ontstond worden besproken, aangezien een van de gevaren is dat de waarheid als zodanig een steeds diffuser begrip wordt door de verwarring die over de authenticiteit van materiaal kan ontstaan. Niet alleen is het probleem dat nepcontent voor echt kan doorgaan. Ook andersom, dat echte beelden voor nep worden versleten kan een significant probleem opleveren.[39] Daarnaast worden voorbeelden besproken van gemanipuleerde content die niet door geavanceerde technieken zijn geproduceerd, die dienen ter illustratie van hoe deepfakes in de toekomst mogelijk zouden kunnen worden ingezet, maar dan op een nog realistischere wijze.

In dit rapport worden de termen deepfake en deepfaketechnologie afwisselend gebruikt. Deepfakes zijn daarbij het product (foto, video, audio, etc.), de technologie dient om de deepfake te genereren. In dit rapport wordt een functionele in plaats van een technische benadering gehanteerd. Het is dus niet zo dat een eindresultaat alleen als deepfake te gelden heeft als het met een bepaalde technologie is vervaardigd of gegenereerd. Nadruk zal liggen op het eindresultaat.

Tot slot nog een beperking van dit rapport. Het onderzoek is uitgevoerd door personen met een juridische achtergrond. De technologie achter deepfakes wordt beschreven in paragraaf 2.1. Daarbij is dicht bij de literatuur en de geciteerde bronnen gebleven. Bovendien zijn de resultaten van de literatuurstudie in de voor deze studie gehouden interviews aan technische experts voorgelegd, om de beschrijving op accuratesse en volledigheid te toetsen. Toch blijft het een beschrijving van de techniek door en voor juristen en beleidsmakers, die waarschijnlijk anders zou zijn aangepakt en ingebed mocht die zijn geschreven door personen met een technische achtergrond. Daarbij is een functionele benadering gekozen en met name gelet op de toepassing(smogelijkheden) van de technische middelen.

1.3 Probleemstelling en onderzoeksvragen

De probleemstelling van dit onderzoek luidt:

> *'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?'*

Aan deze hoofdvraag is een aantal subvragen gekoppeld:

- 1. Welke uitingsvormen van deepfake zijn er te onderscheiden op basis van de bovenstaande indeling?
- 2. Hoe valt het maken en verspreiden van deepfakes binnen de huidige strafrechtelijke bepalingen? Specifiek wordt er gekeken naar de relatie met de volgende delicten:
  - Fraude
  - Afpersing
  - Identiteitsfraude
  - Discriminatie
  - Smaad(schrift), laster en belediging
  - Misbruik van seksueel beeldmateriaal (wraakporno)
  - Gegevensaantasting
- 3. Voldoet het huidige strafrecht om de makers van strafbare deepfakes aan te kunnen pakken?
- 4. Hoe valt het maken van een deepfake video binnen de huidige kaders van het gegevensbeschermingsrecht?
  - Van wie worden er persoonsgegevens verwerkt?
  - Hoe duiden we de verwerking (wat voor verwerking is dit)?
- 5. Welke mogelijkheden hebben burgers op dit moment om deepfakes van het internet te verwijderen en biedt dit hen voldoende handvatten?
- 6. Welke sanctie- of schadevergoeding mogelijkheden zijn er voor het onrechtmatig gebruik van persoonsgegevens voor het maken van deepfake video's?
  - Schadevergoeding vorderen civiel?
  - Verwijderingsgebod?

- 7. Ligt het in de rede dat de Autoriteit Persoonsgegevens hier (ook) als handhaver op gaat treden, of ligt een strafrechtelijke benadering meer voor de hand?
- 8. In hoeverre biedt een vordering uit onrechtmatige daad mogelijkheid om content (die niet onrechtmatig is wegens inbreuk op gegevensbeschermingsecht) van het internet te verwijderen?
- 9. Hoe is het tegengaan van schadelijke of onrechtmatige deepfakes door andere landen gereguleerd voor wat betreft de diverse relevante rechtsgebieden en wat zijn daar eventueel reeds bekende voor- en nadelen van?
- 10. Dienen huidige en toekomstige onrechtmatige of strafbare uitingsvormen van deepfake technologie te leiden tot aanpassingen van de bestaande wetten en regels (met name AVG en strafrecht), of is bestaande wetgeving toereikend?
- 11. Indien aanpassingen worden voorgesteld: in hoeverre zorgen deze aanpassingen voor belemmeringen in de verdere ontwikkeling van bonafide deepfake toepassingen?

De probleemanalyse van dit onderzoek spitst zich toe op vier rechtsgebieden, namelijk het gegevensbeschermingsrecht, het strafrecht, het recht op reputatie/de vrijheid van meningsuiting en het civiel recht, waaronder zowel de onrechtmatige daad als het intellectueel eigendomsrecht wordt bespreken. Ook de AI Act zal kort worden aangestipt, maar die is ten tijde van schrijven nog in ontwikkeling. Daarbij geldt dat het gegevensbeschermingsrecht, het recht op reputatie en vrijheid van meningsuiting en het intellectueel eigendomsrecht rechtsgebieden zijn die zowel bescherming bieden aan burgers die slachtoffer zijn van deepfakes, maar ook ruimte bieden aan anderen die deepfakes maken of verspreiden. Het gegevensbeschermingsrecht heeft bijvoorbeeld twee doelen: het beschermen van de burger (het datasubject) en het faciliteren van (rechts)personen die data willen verwerken (de gegevensverwerkingsverantwoordelijke). Zo ook geeft het regime aangaande de vrijheid van meningsuiting (Artikel 10 EVRM) zowel ruimte aan (rechts)personen die uitingen willen doen, bijvoorbeeld door middel van deepfakes, als aan (rechts)personen die daarin worden afgebeeld of daar anderszins last van kunnen hebben. Het regime aangaande intellectueel eigendom geeft bescherming aan auteurs van een werk, maar biedt tegelijkertijd regels en uitzonderingen voor het gebruik van dat werk door derden.

In deze studie zal de nadruk liggen op de beschermingszijde van deze regimes, alsmede de door het strafrecht en de onrechtmatige daad geboden bescherming, maar zal ook kort worden stilgestaan bij de ruimte die deze rechtsgebieden laten voor het vervaardigen, het verspreiden en het gebruiken van deepfakes, aangezien dit van belang is voor de uiteindelijke conclusies en aanbevelingen van dit rapport. Tot slot is van belang dat alhoewel de hoofdvraag voor dit onderzoek ziet op de negatieve kanten van deepfakes en hoe deze geadresseerd kunnen worden, er ook aandacht zal zijn voor de positieve toepassingen van deepfaketechnologie. Ook die toepassingen zijn immers van belang bij de keuze voor de meest geschikte reguleringsopties.
Hieronder zal kort de problematiek vanuit de vier vermelde rechtsgebieden worden aangestipt:

AVG

Foto's en filmpjes maken en publiceren van anderen mag niet zomaar. Bij de verwerking van persoonsgegevens geldt de Algemene Verordening Gegevensbescherming (AVG)[40] en

de daaraan gekoppelde Uitvoeringswet AVG.[41] Het maakt hierbij wel verschil wie een foto of filmpje maakt en wat diegene daarmee doet. Maakt iemand foto's en filmpjes voor eigen consumptie, dan geldt de AVG in principe niet (dit wordt de 'huishoudelijke exceptie' genoemd). De voorwaarde hierbij is dat deze persoon de foto's en filmpjes privé houdt of hooguit in een zeer beperkte kring deelt, bijvoorbeeld in een kleine appgroep. Zo gauw de foto's en filmpjes in bredere kring worden gedeeld, bijvoorbeeld op een openbare Facebookpagina, dan geldt de huishoudelijke exceptie niet meer. Omdat de AVG dan van toepassing is, is er een legitieme verwerkingsgrondslag nodig, zoals dat er toestemming is gegeven door de personen die op de foto's en filmpjes staan of dat het belang dat met de verspreiding van deepfakes wordt gediend groter is dan het belang van die personen om hun gegevens niet verwerkt te zien.

Het gebruik van foto's of filmmateriaal dat reeds online staat (bijvoorbeeld voor het maken van deepfake video's) is niet zomaar toegestaan. Dat deze beelden al op internet staan, betekent niet dat iemand ze zomaar mag gebruiken voor nieuwe toepassingen. De gegevens worden dan namelijk in een andere context gebruikt en voor een ander doel, wat volgens de AVG slechts onder voorwaarden is toegestaan. Staan iemands gegevens op het internet, dan heeft diegene het recht om te verzoeken deze gegevens te laten verwijderen, bijvoorbeeld omdat ze niet langer nodig zijn voor het oorspronkelijke doel waarvoor ze verzameld of verwerkt zijn, er geen rechtsgrond (meer) is voor verwerking of omdat de verwerking onrechtmatig is. Omdat het fenomeen deepfake veel verschijningsvormen kent en deepfakes bovendien met verschillende intenties wordt gemaakt, (soms is het grappig en duidelijk niet bedoeld om een persoon te beschadigen) is het zinvol om middels dit onderzoek na te gaan hoe deze zich verhouden tot de AVG.

Strafrecht

Het maken en verspreiden van nepvideo's komt als delict in het strafrecht, zoals gegeven door het Wetboek van Strafrecht (Sr), niet voor. Toch zijn er vanuit de opsporing zorgen geuit. Zo maakte het Openbaar Ministerie zich zorgen om het feit dat de deepfaketechnologie als middel voor afpersing kan gaan dienen, al ziet het niet direct aanleiding om het bestaande strafrecht aan te passen.[42] Veel kwalijke uitingsvormen kunnen reeds met het huidige strafrecht worden aangepakt, zoals met gebruikmaking van bestaande delictsomschrijvingen, onder meer aangaande chantage, afpersing en oplichting. Toch is het belangrijk om na te gaan of nieuwe strafbaarstellingen van deepfaketoepassingen gewenst zijn, waarbij met name kan worden gedacht aan de strafbaarstelling van (deepfake) wraakporno (waarbij uiteraard geldt dat 'wraak' in deze context wat misplaatst is omdat het slachtoffer doorgaans niet kwaads heeft gedaan).

Op 1 januari 2020 is er een wet[43] in werking getreden waarin wraakporno op zichzelf strafbaar wordt gesteld. Deze wet houdt in dat het verboden is om seksueel getinte afbeeldingen van anderen te delen wanneer de dader wist dat het delen ervan nadelig zou zijn voor het slachtoffer. Of het slachtoffer toestemming heeft gegeven voor het maken van de afbeeldingen is daarbij niet van belang. In de praktijk betekent dit dat het delen van naakt- of seksueel getinte afbeeldingen sneller bestraft kan worden. Interessant is om na te gaan hoe de strafbaarstelling van wraakporno zich verhoudt tot deepfakes van seksuele aard

en welke vormen van niet-seksuele inhoud reeds onder het strafrecht zijn geadresseerd en hoe.

Vrijheid van meningsuiting en recht op reputatie

Vervolgens is het belang dat binnen de Nederlandse en de Europese rechtsorde de vrijheid van meningsuiting geldt. Zo stelt artikel 10 van het Europees Verdrag voor de Rechten van de Mens (EVRM):[44]

1. Een ieder heeft recht op vrijheid van meningsuiting. Dit recht omvat de vrijheid een mening te koesteren en de vrijheid om inlichtingen of denkbeelden te ontvangen of te verstrekken, zonder inmenging van enig openbaar gezag en ongeacht grenzen. Dit artikel belet Staten niet radio- omroep-, bioscoop of televisieondernemingen te onderwerpen aan een systeem van vergunningen.
2. Daar de uitoefening van deze vrijheden plichten en verantwoordelijkheden met zich brengt, kan zij worden onderworpen aan bepaalde formaliteiten, voorwaarden, beperkingen of sancties, die bij de wet zijn voorzien en die in een democratische samenleving noodzakelijk zijn in het belang van de nationale veiligheid, territoriale integriteit of openbare veiligheid, het voorkomen van wanordelijkheden en strafbare feiten, de bescherming van de gezondheid of de goede zeden, de bescherming van de goede naam of de rechten van anderen, om de verspreiding van vertrouwelijke mededelingen te voorkomen of om het gezag en de onpartijdigheid van de rechterlijke macht te waarborgen.

Deze vrijheid is zeer ruim en omvat, zoals dat in het Engels heet, 'the right to offend, shock and disturb'. Het verspreiden van informatie die leugenachtig is, is niet verboden, tenzij een van de belangen als genoemd in lid 2 van artikel 10 EVRM wordt geraakt, zoals de nationale veiligheid of de eer een goede naam van personen. Bij beide gronden heeft het Europees Hof voor de Rechten van de Mens (EHRM) de lat hoog gelegd. Zo heeft het Hof gesteld dat publieke personen, zoals royalty en politici, meer moeten accepteren in termen van satire, kritiek en het openbaar maken van privégegevens dan anderen. Wel kunnen zij zich beroepen op hun recht op privéleven, zoals vervat in artikel 8 EVRM. Daarom is het belangrijk te onderzoeken waar de grens precies ligt en welk type deepfake onder welke omstandigheden een mensenrechtenschending zou betekenen.

Onrechtmatige Daad en Intellectueel Eigendom

Tot slot is van belang het Nederlands civiel recht en dan met name twee deelgebieden.

Enerzijds kunnen burgers die ongewenst zijn geportretteerd in een deepfake zich beroepen op hun persoonlijkheidsrecht, zoals met name volgt uit artikel 21 van de Auteurswet:[45]

Is een portret vervaardigd zonder daartoe strekkende opdracht, den maker door of vanwege den geportretteerde, of te diens behoeve, gegeven, dan is openbaarmaking daarvan door dengene, wien het auteursrecht daarop toekomt, niet geoorloofd, voor zoover een redelijk belang van den geportretteerde of, na zijn overlijden, van een zijner nabestaanden zich tegen de openbaarmaking verzet.

Onderzoeksvragen die hierbij naar voren komen zijn onder meer: in hoeverre rust er auteursrecht op een samenvoeging van beelden of op nieuw gegenereerde content op basis van bestaand materiaal, in welke mate gelden hier

morele en persoonlijkheidsrechten, in hoeverre kunnen overledenen, of hun directe naasten namens hen, een beroep doen op intellectueel eigendomsrechten en in hoeverre vallen deepfake audiofragmenten ook onder deze doctrine? Het is van belang het intellectueel eigendomsrecht te bestuderen, omdat dit rechtsgebied bescherming biedt aan de maker van een werk dat wordt gebruikt voor de vervaardiging van gemanipuleerd materiaal, ruimte biedt aan het hergebruik van materiaal onder bepaalde voorwaarden, bescherming biedt aan de geportretteerde en regels bevat aangaande het downloaden en gebruiken van bestaand of gemanipuleerd materiaal.

Anderzijds kan de publicatie of verspreiding van een deepfake een onrechtmatige daad opleveren in de zin van artikel 6:162 Burgerlijk Wetboek (BW), dat luidt:

1. Hij die jegens een ander een onrechtmatige daad pleegt, welke hem kan worden toegerekend, is verplicht de schade die de ander dientengevolge lijdt, te vergoeden.
2. Als onrechtmatige daad worden aangemerkt een inbreuk op een recht en een doen of nalaten in strijd met een wettelijke plicht of met hetgeen volgens ongeschreven recht in het maatschappelijk verkeer betaamt, een en ander behoudens de aanwezigheid van een rechtvaardigingsgrond.
3. Een onrechtmatige daad kan aan de dader worden toegerekend, indien zij te wijten is aan zijn schuld of aan een oorzaak welke krachtens de wet of de in het verkeer geldende opvattingen voor zijn rekening komt.

Burgers die stellen schade te hebben ondervonden van een deepfake kunnen de rechter vragen de maker of verspreider van dat bericht te gebieden deze schade te vergoeden en de verdere verspreiding te staken. Dat kan bijvoorbeeld zijn als een deepfake in strijd is met het recht, zoals de AVG, of als er zorgvuldigheidsnormen worden overschreden. Daarom zal worden onderzocht in hoeverre het leerstuk van onrechtmatige daad toereikend is voor deepfake berichten.

1.4 Methoden

Dit rapport betreft een verkennend onderzoek. In plaats van een diepgravende analyse op specifieke onderdelen, richt het zich op een verkennende, brede inventarisatie van mogelijke lacunes of onvolkomenheden in de wetgeving en van reguleringsopties om deze lacunes of onvolkomenheden te adresseren. Voor deze verkenning zijn literatuuronderzoek en juridische analyse de belangrijkste methoden, die worden aangevuld met interviews en een landeninventarisatie.

Voor het *literatuuronderzoek* is een breed scala aan literatuur bestudeerd, niet alleen wetenschappelijke bronnen maar ook niet-wetenschappelijke bronnen, aangezien deepfake pas sinds relatief kort in de belangstelling staat en daarom nog niet uitgebreid in juridische zin is geanalyseerd. Ook de snel voortschrijdende technologie maakt het relevant om webpagina's en nieuwsberichten te bestuderen die nieuwe ontwikkelingen reflecteren. Bij het literatuuronderzoek is vooral literatuur over deepfaketechnologie en toepassingen bestudeerd, evenals literatuur over de maatschappelijke gevolgen daarvan.

Voor de *juridische analyse* is de klassiek-juridische methode van doctrinaire wetsanalyse

gebruikt. Deze bestaat uit het analyseren of, hoe en in welke mate de bestaande en aanhangige wetgeving van toepassing is op een nieuw fenomeen, in dit geval deepfakes. Hierbij wordt de Nederlandse wetgeving – vanzelfsprekend inclusief de daarbij geldende Europeesrechtelijke kaders en regelgeving – geanalyseerd op basis van jurisprudentie-analyse en grammaticale, wetshistorische en teleologische wetsinterpretaties. Door te analyseren hoe wetsbepalingen en doctrines van toepassing zijn op deepfakes, worden mogelijke lacunes of onvolkomenheden in de wet blootgelegd die mogelijk afbreuk doen aan de rechtsbescherming van burgers tegen kwalijke gevolgen van deepfakes.

De bevindingen uit het literatuuronderzoek en de juridische analyse worden verdiept en aangevuld met vijftien interviews met experts. Deze experts zijn geselecteerd uit binnen- en buitenland en uiteenlopende expertises (zie par. 6.4 voor een nadere toelichting op de selectie). De interviews zijn verricht met een semigestructureerde vragenlijst om de eerste onderzoeksresultaten te valideren en verdere verbreding en verdieping te vinden. Daarbij is met name gekeken naar de effecten van deepfakes – welke deepfakes hebben de grootste impact, welke groepen worden met name geraakt door deepfakes, enzovoorts – en naar mogelijke reguleringsoplossingen, met name aanpassingen in het juridische kader. Ook is gevraagd hoe deepfaketechnologie zich naar verwachting in de komende jaren zal ontwikkelen. Om ervoor te zorgen dat reguleringsopties niet reeds achterhaald zijn op het moment dat ze worden ingevoerd, is ook gevraagd welke karakteristieken regulering volgens de experts dient te hebben om toekomstbestendig te zijn.

Tot slot is een *landeninventarisatie* gemaakt om de bevindingen te verdiepen en aan te vullen, met name om een bredere blik te krijgen op reguleringsmogelijkheden. Op basis van een quickscan op internet van buitenlandse wetgeving en wetsvoorstellen die deepfakes reguleren, zijn elf landen geïdentificeerd met potentieel interessante wetgevingsinitiatieven. Van deze landen bleken twee landen de meest interessante verdergaande regulering te hebben op het gebied van deepfake, namelijk China en de Verenigde Staten (zowel op federaal als op statelijk niveau). Voor deze landen zijn landenrapporteurs gevraagd overzichten te schetsen van deze regulering, die zijn gebruikt om de voorlopige bevindingen ten aanzien van reguleringsopties uit te breiden of aan te scherpen.

1.5 Aanpak en leeswijzer

De leeswijzer is als volgt:

- Hoofdstuk 2 beschrijft de deepfaketechnologie, schetst de diverse toepassingsmogelijkheden, de positieve en negatieve effecten daarvan en geeft een typologie van deepfakes en de toepassingsmogelijkheden die zal dienen als basis voor de juridische analyse.
- Hoofdstuk 3 geeft een beschrijving van het huidige regulerende kader ten aanzien van deepfakes en analyseert in hoeverre het vigerende materieelrecht toereikend is en waar mogelijke lacunes bestaan. Hierbij zal worden stilgestaan bij de vier eerdergenoemde rechtsgebieden, te weten het strafrecht, het gegevensbeschermingsrecht, de vrijheid van meningsuiting en het civiel recht.
- Hoofdstuk 4 gaat in op bewijslastkwesties. Welke invloed heeft de introductie van deepfakes op rechtszaken, waar ligt de bewijslast voor de echtheid van materiaal en waar zou die moeten komen te liggen en hoe hoog ligt de lat voor

toelaatbaar bewijs? Zijn daarin verschillen voor de beantwoording van deze vragen te ontwaren tussen de diverse rechtsgebieden, zijn er lacunes in het huidige rechtssysteem en hoe zouden die eventueel kunnen worden opgevuld?

- Hoofdstuk 5 gaat in op handhavingsvragen en op dilemma's rond het toezicht en de handhaving van de verschillende materieelrechtelijke bepalingen. Dit vraagstuk zal worden besproken in het bredere licht van rechtshandhaving in horizontale privacyrelaties als zodanig, omdat veel van de problematiek in dit kader voor deepfakes gelijk is aan de vragen die spelen rond de verwerking van persoonsgegevens door burgers door middel van drones, smartphone, gezichtsherkenningsapps of spionageapparatuur.
- Hoofdstuk 6 bespreekt de regulering van deepfakes in het buitenland. Hierbij zal een algemene quickscan worden gegeven van de verschillendediscussiesenreguleringsstrategieën over de wereld. Specifieke landenrapportages, die als bijlage bij dit rapport zullen worden gevoegd, zullen worden geschreven door twee externe experts ten aanzien van twee landen die het verst zijn met de regulering van deepfakes: de Verenigde Staten en China. De belangrijkste lessen uit die rapportages voor de Nederlandse context zullen in dit hoofdstuk aan bod komen. De volledige rapportages staan in de bijlagen bij dit rapport. Ook zal een aantal interviews worden gehouden met internationale experts. De gespreksverslagen daarvan zijn te vinden in de bijlagen bij dit rapport en de belangrijkste bevindingen daaruit zullen in dit hoofdstuk worden weergegeven.
- Hoofdstuk 7 geeft een eerste voorlopige samenvatting en de reflecties daarop.
- Hoofdstuk 8 biedt een schets van de bestaande juridische lacunes en geeft een aantal reguleringsopties om deze lacunes te adresseren.
- Hoofdstuk 9 bevat de conclusie van dit onderzoek en staat kort stil bij de beantwoording van de hoofdvraag en de diverse deelvragen die voor dit onderzoek gegeven zijn.
- Hoofdstuk 10 bevat de bijlagen, waarin zijn vervat twee landenstudies, namelijk die ten aanzien van China en de Verenigde Staten, en de volledige interviewverslagen.

Voetnoten hoofdstuk 1

- 1 <https://www.youtube.com/watch?v=PpwHIGHBfuk>.
- 2 <https://www.youtube.com/watch?v=IvY-Abd2FfM>.
- 3 <https://www.dailystar.co.uk/news/world-news/elderly-women-chinese-tiktok-scammed-22971322>.
- 4 <https://www.inputmag.com/tech/open-source-program-will-let-you-run-deepfakes-on-live-video-calls>.
- 5 <https://www.bbc.com/news/business-56278411>. <https://www.nytimes.com/2019/11/24/technology/tech-companies-deepfakes.html>.
- 6 Trump heeft laten zien dat zelfs zonder deepfakes het mogelijk is om zo'n driekwart van zijn kiezers in een of meer leugens over zijn verkiezingsnederlaag te laten geloven. <https://edition.cnn.com/2021/02/04/politics/2020-election-donald-trump-voter-fraud/index.html>.
- 7 <https://www.bnr.nl/nieuws/internationaal/10338611/vertrouwen-in-media-was-nooit-eerder-zo-laag>. <https://www.edelman.com/trust/2021-trust-barometer/press-release>.

- 8 Custers, B. (2019). Nepnieuws, filterbubbels en echokamers. *In: S. van der Hof, BHM Custers, F. Dechesne en ELO Keymolen (eds.) Recht uit het hart, Universiteit Leiden: Meijers Instituut.*
- 9 <https://io9.gizmodo.com/why-reality-fatigue-has-made-science-fiction-more-int-346941>.
- 10 Media zijn in de loop der jaren steeds afhankelijker zijn geworden van content door burgers aangeleverd. Esneijer, J. et. al. (2012). Making user created news work. TNO report, (R11277). Shah, R., & Zimmermann, R. (2017). *Multimodal analysis of user-generated multimedia content.* Springer International Publishing.
- 11 Parchomovsky, G. (2000). Publish or perish. Michigan Law Review, 98(4), 926-952.
- 12 <https://www.euronews.com/2018/11/22/france-passes-controversial-fake-news-law>.
- 13 < https://www.volkskrant.nl/cultuur-media/regisseur-hans-block-over-zijn-onthullende-documentaire-the-cleaners-waarin-de-schaduwzijde-van-facebook-wordt-belicht~b042f501/>.
- 14 Maras, M. H., & Alexandrou, A. (2019). Determining authenticity of video evidence in the age of artificial intelligence and in the wake of Deepfake videos. *The International Journal of Evidence & Proof, 23(3).*
- 15 <https://www.telegraph.co.uk/news/2020/01/31/deepfake-audio-used-custody-battle-lawyerreveals-doctored-evidence/>.
- 16 <https://www.forbes.com/sites/jessedamiani/2019/09/03/a-voice-deepfake-was-used-to-scam-a-ceo-out-of-243000/#5b10b18022241>.
- 17 Zie meer in het algemeen het zeer rijke rapport van Interpol: <https://www.europol.europa.eu/publications-documents/malicious-uses-and-abuses-of-artificial-intelligence>.
- 18 <https://www.bbc.com/news/technology-48621452>.
- 19 <https://www.malaymail.com/news/malaysia/2019/06/12/is-the-political-aide-viral-sex-video-confession-real-or-adeepfake/1761422>. <https://apnews.com/4841d0ebcc704524a38b1c8e213764d0>.
- 20 Jankowicz, N. (2020). *How to Lose the Information War: Russia, Fake News, and the Future of Conflict. IB Tauris, Limited.*
- 21 <https://edition.cnn.com/interactive/2019/01/business/pentagons-race-against-deepfakes/>
- 22 Tweede Kamer, Brief heimelijke beïnvloeding van de publieke opinie door statelijke actoren, 8e1652c7-or1-3.0. <https://www.aivd.nl/actueel/nieuws/2017/12/18/buitenlandse-beinvloeding-bestrijden-door-samenwerking-overheden-media-en-bedrijfsleven>.
- 23 Bunn, A. (2019). Children and the 'Right to be Forgotten': What the right to erasure means for European children, and why Australian children should be afforded *a similar right.* Media International Australia, 170(1).
- 24 'Slut-shaming is een Engelstalige term die verwijst naar het bestempelen van een meisje of vrouw als slet of hoer omwille van haar seksueel gedrag. Het gaat dan vaak over wat gepercipieerd wordt als seksueel gedrag. Met de opkomst van nieuwe media worden deze termen onder jongeren ook op sociale media gebruikt. Hoewel vrienden onder elkaar woorden zoals 'slet' en 'hoer' gebruiken zonder kwetsende bedoeling, neemt het toch vaak de vorm van beledigingen of pesten aan.' <https://www.stoppestennu.nl/slut-shaming-op-social-media-bangalijsten>.
- 25 <https://mediawijs.be/dossiers/dossier-cyberpesten/bijna-1-5-tienermeisjes-krijgt-te-maken-slut-shaming-sociale-media>; <https://www.aauw.org/app/uploads/2020/03/Crossing-the-Line-Sexual-Harassment-at-School.pdf>.
- 26 Van Royen, K., Poels, K., Vandebosch, H., & Walrave, M. (2018). Slut-Shaming 2.0. In Sexting (pp. 81-98). Palgrave Macmillan, Cham.
- 27 Maddocks, S. (2020). 'A Deepfake Porn Plot Intended to Silence Me': exploring continuities between pornographic and 'political'deep fakes. Porn Studies, 7(4), 415-423.
- 28 Westerlund, M. (2019). The emergence of deepfake technology: A review. *Technology Innovation Management Review, 9*(11).
- 29 <https://www.bbc.com/news/technology-56404038>.

- 30 Lester, D., McSwain, S., & Gunn III, J. F. (2013). Suicide and the Internet: the case of Amanda Todd. *International Journal of Emergency Mental Health and Human Resilience*.
- 31 Keats Citron, D. (2018). Sexual privacy. Yale LJ, 128, 1870.
- 32 Tweede Kamer, vergaderjaar 2018–2019, 34 926, nr. 8.
- 33 <https://www.congress.gov/bill/116th-congress/senate-bill/2065>.
- 34 <https://www.theverge.com/2019/11/29/20988363/china-deepfakes-ban-internet-rules-fake-news-disclosure-virtual-reality>.
- 35 <https://www.europol.europa.eu/newsroom/news/new-report-finds-criminals-leverage-ai-for-malicious-use-%E2%80%93-and-it%E2%80%99s-not-just-deep-fakes>.
- 36 <https://www.un.org/en/chronicle/article/how-can-multilateralism-survive-era-artificial-intelligence>. <https://indiaai.gov.in/news/india-spotlights-the-dangers-of-deepfakes-for-world-peace-at-unsc>.
- 37 <https://www.euractiv.com/section/cybersecurity/news/estonian-intelligence-russians-will-develop-deep-fake-threats/>.
- 38 Tweede Kamer, vergaderjaar 2018–2019, 34 926, nr. 2.
- 39 Zie o.a. de verwarring die ontstond: <https://www.nrc.nl/nieuws/2021/04/23/door-schrijver-gepubliceerde-en-verwijderde-video-van-rutte-blijkt-niet-nep-a4041126?s=09>.
- 40 Verordening (EU) 2016/679 van het Europees Parlement en de Raad van 27 april 2016 betreffende de bescherming van natuurlijke personen in verband met de verwerking van persoonsgegevens en betreffende het vrije verkeer van die gegevens en tot intrekking van Richtlijn 95/46/EG (algemene verordening gegevensbescherming)
- 41 Uitvoeringswet Algemene verordening gegevensbescherming.
- 42 <https://nos.nl/artikel/2300688-zorgen-om-over-deepfakes-risico-op-oplichting-en-afpersing.html>.
- 43 Wijziging van onder meer het Wetboek van Strafrecht in verband met de herwaardering van de strafbaarstelling van enkele actuele delictsvormen (herwaardering strafbaarstelling actuele delictsvormen), Stb. 2019, 311.
- 44 Verdrag tot Bescherming van de Rechten van de Mens en de Fundamentele Vrijheden Rome, 4.XI.1950.
- 45 Wet van 23 september 1912, houdende nieuwe regeling van het auteursrecht.

Een deepfake kan worden gegenereerd op basis van bestaande content. Bv: duizend foto's worden gebruikt om een nieuwe (fake) foto te genereren die nauwelijks van echt te onderscheiden is.

2. Deepfakes

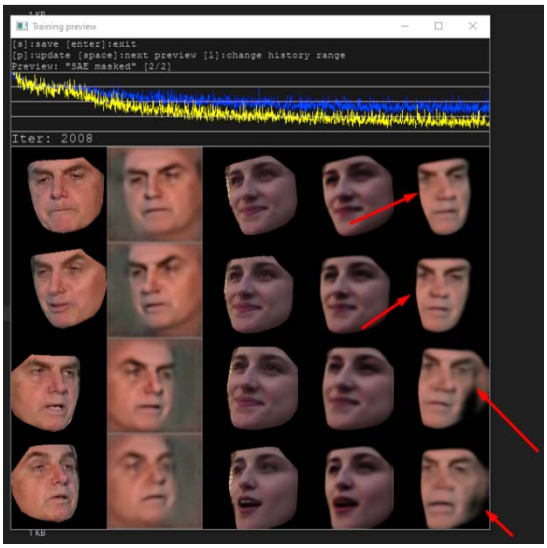
Dit hoofdstuk zal een korte schets geven van de ontwikkeling en technische achtergrond van deepfaketechnologiën (paragraaf 2.1), een overzicht van de diverse positieve (paragraaf 2.2) en negatieve (paragraaf 2.3) consequenties van deepfakes bieden, een bespreking van de grotere maatschappelijke effecten (paragraaf 2.4) en een typologie geven (paragraaf 2.5), die dient als uitgangspunt voor de juridische analyse die in de volgende hoofdstukken aan bod komt. Tot slot wordt een korte conclusie gegeven (paragraaf 2.6). De schets die in dit hoofdstuk wordt geboden betreft het fenomeen deepfake in de breedte: het geeft een kort overzicht van de meeste toepassingen van deepfakes zoals die nu bekend of voorzienbaar zijn. In hoofdstuk 3 wordt vervolgens ingegaan op een deel van deze toepassingen, namelijk voor zover zij in horizontale relaties (dat wil zeggen: tussen burgers onderling) worden ingezet. Voor de juridische analyse en regulering van de inzet van deepfakes door bedrijven of statelijke actoren of door burgers jegens bedrijven en statelijke actoren zal vervolgonderzoek nodig zijn. In dit hoofdstuk is ervoor gekozen om de maatschappelijke effecten van deepfakes in een aparte paragraaf te bespreken; het gaat hierbij om gevolgen die niet beperkt zijn tot de inzet van specifieke deepfakes in specifieke gevallen, maar om systemische effecten die gerelateerd zijn aan het bestaan en het gebruik van deepfakes als zodanig.

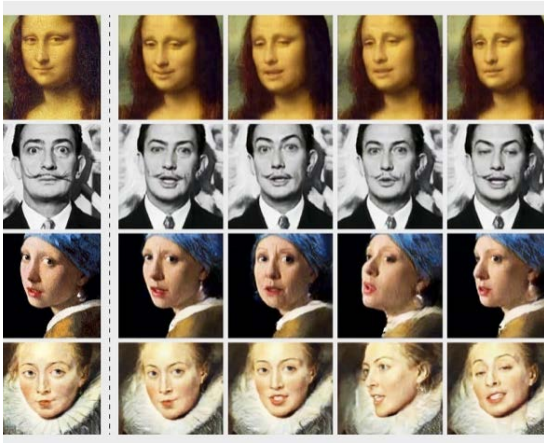
2.1 Historie en technologische achtergrond

Er zijn momenteel al tientallen apps op de markt die burgers kunnen downloaden, bijvoorbeeld om zichzelf te laten figureren in een bekende Hollywoodfilm, om woorden in de mond van politici of vrienden te leggen, een geheel eigen versie van het NOS-journaal te genereren, historische figuren tot leven te wekken[46] en bekenden handelingen te laten verrichten die zij niet daadwerkelijk hebben verricht.[47] Gezichtsuitdrukkingen kunnen worden gemanipuleerd per frame, toonhoogte, timbre en taalgebruik kunnen al naar gelang de wensen van de maker worden aangepast en identiteiten van twee of meer personen kunnen worden samengevoegd, bijvoorbeeld door de gezichten van twee personen samen te laten smelten of door een karakter het aanzicht van een bekend persoon te geven en het stemgeluid van een ander, waarbij de mondbewegingen van de een naadloos aansluiten op de tekst gesproken met het stemgeluid van de ander.[48] Deepfaketechnologie bestaat pas een aantal jaren; de kwaliteit en snelheid waarmee ze kunnen worden gegenereerd is sindsdien met rassenschrede vooruit gegaan.

Deepfake is de benaming van een Machine Learning (ML)-technologie, waarbij gebruik wordt gemaakt van Artificial Intelligence (AI) en deep learning. Met behulp van Artificial Neural Networks (ANN), die zijn gebaseerd op 'echte' oftewel biologische neurale netwerken waarmee dieren zijn behept, kunnen systemen leren hoe zij taken moeten uitvoeren door naar voorbeelden te kijken, zonder dat er specifieke regels in het systeem worden geprogrammeerd.[49] 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)[50] en Variational AutoEncoders

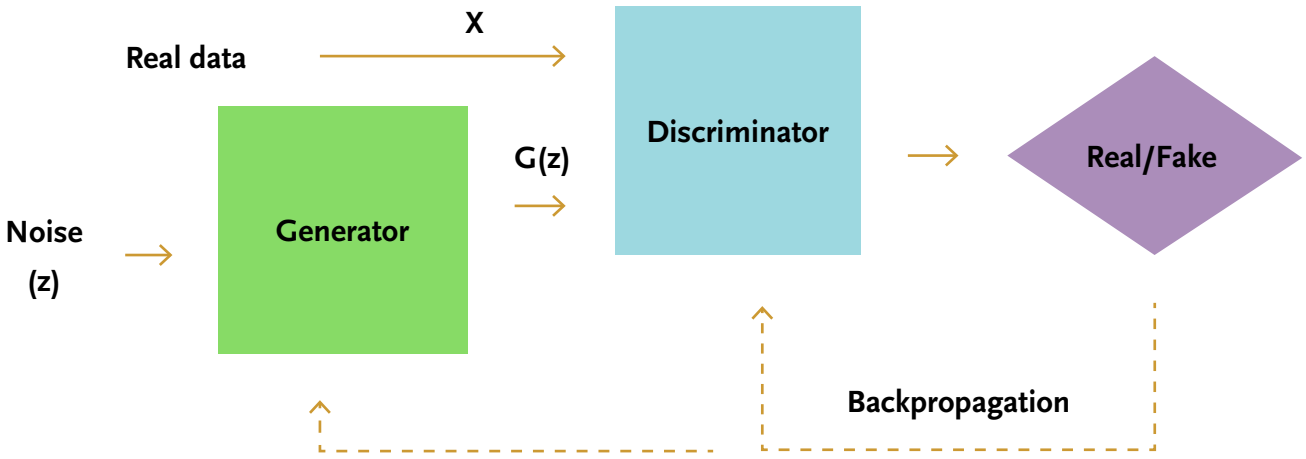
(VAE). [51] GAN is een techniek die is uitgevonden door Ian Goodfellow in 2014. Deze techniek heeft de grenzen van de moderne resultaten verlegd en de kwaliteit en de resolutie van de geproduceerde beelden verbeterd en loopt voorop als het gaat om betrouwbaar, met lage kosten en tijdsinvesteringen diverse beelden en geluiden te genereren met modellen die rechtstreeks leren uit bestaande data.



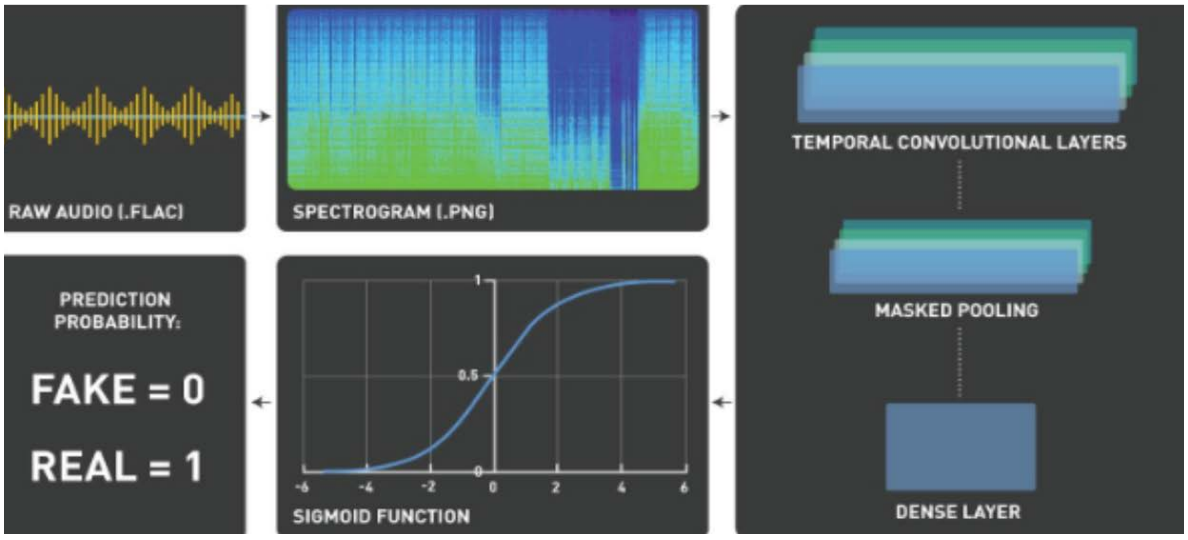


Figuur 2-5: voorbeelden van deepfakes van celebrities [52]

GAN's bestaan uit twee concurrerende netwerken, een generator G (x) en een discriminator D (x). Hierbij worden tegelijkertijd twee concurrerende modellen getraind, om zodoende beelden te synthetiseren.[53] De algoritmen weerspiegelen de inhoud en produceren realistische nepbeelden. Hierbij is het doel om willekeurige ruis toe te wijzen aan samples en om echte en gegenereerde samples te onderscheiden.[54] De twee concurrerende netwerken G en D spelen beide een vijandig spel waarbij de generator de discriminator voor de gek probeert te houden door gegevens te genereren die te vergelijken zijn met die in de bestaande trainingsset. Vervolgens probeert de discriminator zich niet voor de gek te laten houden door nepgegevens te identificeren op basis van echte gegevens. De generator en de discriminator werken tegelijkertijd om complexe taken en output te leren genereren en te herkennen.[55]



Figuur 6: Verbeelding van GAN[56]



Figuur 7: Deepfake audiofragmenten[57]

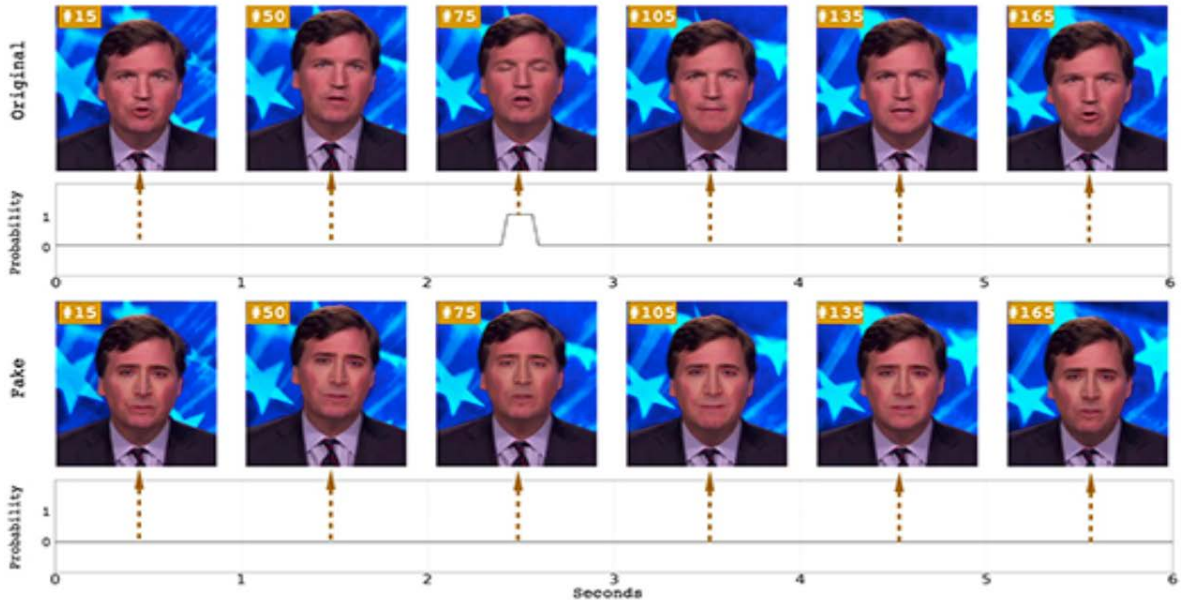
Dit gaat als volgt te werk. De generator genereert afbeeldingen uit willekeurige ruis (z) en leert vervolgens zelf hoe realistische afbeeldingen kunnen worden gegenereerd. De ingevoerde ruis wordt bemonsterd met behulp van uniforme of normale distributie, waarna het naar de generator wordt gevoerd. Deze zal vervolgens een beeld genereren. Deze output, die bestaat uit nepbeelden, en de echte beelden van de trainingsset worden ingevoerd in de discriminator. Die leert vervolgens hoe je nepbeelden van echte beelden kunt onderscheiden. De output D (x) is de kans dat de input echt is, wanneer de input echt is zal D (x) 1 zijn en als het wordt gegeneerd door de generator dan zal D (x) 0 zijn.[58]

Met deze techniek kan door middel van het bekijken van bijvoorbeeld duizend foto's van Donald Trump een nieuwe foto 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.[59] Onderzoekers van Carnegie Mellon University hebben in 2018 een nieuwe techniek ontwikkeld waarbij deepfakes automatisch worden gegenereerd zonder

menselijke tussenkomst (unsupervised),[60] waardoor bijvoorbeeld *'Barack Obama's style can be transformed into Donald Trump".*[61]

Hoewel de techniek aanvankelijk complex was, significante kosten met zich meebracht en gespecialiseerde apparatuur en programma's vereiste, is het nu voor iedereen mogelijk om een deepfake te creëren in een handomdraai.[62] Een smartphone of computer is voldoende. Deepfakes kunnen reeds worden gefabriceerd uit een enkel stilstaand beeld[63] en de bewegingen van het hele lichaam kunnen daarbij worden meegenomen.[64] Een Japans bedrijf, DataGrid, is er zelfs in geslaagd afbeeldingen te maken van het hele lichaam in hoge resolutie van niet-bestaande mensen.65 Virtuele avatars en deepfake-personages lopen hierdoor in elkaar over.

Sinds het ontstaan van de techniek is er gewerkt aan automatische deepfake-detectiesystemen.[66] Aanvankelijk was dat nog redelijk eenvoudig, bijvoorbeeld omdat deepfaketechnologieën de ogen van personen niet of nauwelijks lieten knipperen.[67] Alhoewel dat inmiddels verholpen is, zijn deepfakes ook nu nog niet perfect, mede omdat ze vaak artefacten bevatten, dat wil zeggen, sporen waaruit de beeld- of geluidsmanipulatie kan blijken. Zo laat elk apparaat een specifiek merkteken achter op afbeeldingen, ook wel een PRNU-patroon genoemd. Dit komt door onvolkomenheden in het fabricageproces van het apparaat.[68] Het PRNU-patroon kan als ware worden beschouwd als een vingerafdruk van een apparaat en kan worden gebruikt om beeldattributie uit te voeren, beeldmanipulaties te detecteren en te lokaliseren.[69] Het ontbreken van zo'n PRNU-patroon kan worden gezien als een aanwijzing voor manipulatie.[70] Hoewel GAN-gegenereerde afbeeldingen niet dezelfde markeringen zullen vertonen, zijn zij wel het resultaat van complexe verwerkingssystemen met een groot aantal filterprocessen, die mogelijk ook sporen achterlaten op outputbeelden.



Figuur 8: 'Professor of Computer Science Siwei Lyu is leading research efforts to use eye-blinking on original video (top) and a DeepFake video (bottom) to determine if a video was counterfeit. Lyu is working with Facebook to improve detection efforts. (Graphic by Yuezun Li, Ming-Ching Chang and Siwei Lyu).'[71]

Tegenwoordig worden vaak deep learning methoden ingezet om deepfakes te ontdekken[72] en kan nog vaak een discrepantie tussen de lipbeweging en de audio worden waargenomen[73] en zijn sommige hoofdhoudingen onnatuurlijk.[74] Hoewel deze detectiemethoden nu nog behulpzaam zijn, zullen deepfaketechnologieën daar op termijn ook weer van leren, door deze kenmerken op te nemen in de discriminator en kan de generator zodoende worden verfijnd om bepaalde maatregelen te treffen om deze kleine foutjes te voorkomen.[75] Vaak vereist het detecteren van deepfakes een grote database waarin zich zowel echte als onechte video's bevinden, om zo de modellen te kunnen trainen. Deepfakes gegenereerd door middel van GAN-modellen zijn over het algemeen moeilijk detecteerbaar, aangezien op basis van bestaande invoergegevens nieuwe output kan worden geproduceerd dat van vergelijkbare kwaliteit is, die alleen niet kan worden vergeleken met bestaand materiaal of residuen daaruit bevat.[76]

In 2019 bestond er nog geen goede dataset om deepfakes te detecteren. Om die reden werkten Facebook, Microsoft, the Partnership on AI coalition en academici van zeven universiteiten[77] samen aan de Deepfake Detection Challenge (DFDC), waarbij het doel was om technologie te ontwikkelen die iedereen kan gebruiken om deepfakes te detecteren. Het beste model zou aanvankelijk een nauwkeurigheid van 82,56% hebben, maar nadat het nogmaals werd getest middels de zogenoemde black box-dataset, daalde het percentage naar 65,18%.[78] Microsoft heeft in 2020 een *Video Authenticator Tool* ontwikkeld,[79] die een foto of een video een betrouwbaarheidsscore kan geven, gerelateerd aan de kans dat de video kunstmatig is. Bij video's kan ten aanzien van elk frame zo'n percentage worden gegeven. Het systeem werkt door de zogenaamde menggrens van de deepfake en de subtiele vervagende of grijswaardenelementen te detecteren.[80] Microsoft heeft ook een andere technologie ontwikkeld die zowel gemanipuleerde inhoud kan detecteren als dat het authentieke beelden kan verifiëren, onder meer doordat aan digitale content certificaten en hashes, dat wil zeggen een unieke code die als metadata aan de content verbonden is, kunnen worden toegevoegd. Die certificaten en hashes kunnen vervolgens worden geverifieerd als bepaalde content opduikt.[81] De nieuwste tool is in 2021 beladen met prijzen, onder meer omdat die 10 keer sneller is dan veel vorige methoden en bestaat uit zo'n 150.000 coderegels.[82]

De dynamiek tussen deepfakes en deepfake-detection zal een kat-en-muis-spel blijven, waarbij beide steeds verder ontwikkelen en van elkaar leren, net zoals bijvoorbeeld bij encryptie- en decryptiemethoden het geval is. Daarom zal er, of er nu wordt gekozen voor een volautomatische detectieprocedure, een handmatige procedure of een combinatie van beide, altijd onzekerheid blijven over de echtheid van beeld- en geluidfragmenten. Detectiemethoden kunnen een kans geven dat een audio of videobericht echt is.[83] Een vraag die zal moeten worden beantwoord is of alle content op termijn aan een dergelijke waarheidscheck zal moeten worden onderworpen of dat dit alleen hoeft binnen speciale contexten of bij reden tot twijfel. Ook is van belang te kijken naar welke 'waarheidspercentages' of 'betrouwbaarheidspercentages' er moeten worden gehanteerd en of dit voor diverse toepassingsgebieden verschillend moet zijn. Dient de overheid een hoger waarheidspercentage als standaard te hanteren

dan het bedrijfsleven, moet binnen het strafrecht het waarheidspercentage hoger zijn dan binnen het civielrecht, moet de NOS een hoger waarheidspercentage hanteren dan commerciële nieuwsorganisaties?

2.2 Positieve toepassingen (kansen)

Deepfakes bieden een flink aantal kansen en risico's, waarbij in ogenschouw moet worden genomen dat een kans of positieve toepassing voor de een niet zelden een risico of negatieve toepassing voor de ander betekent. Onderstaand overzicht is indicatief, mede vanwege de grote hoeveelheid toepassingen en de diversiteit daarvan; gezien het feit dat de technologie zich in hoog tempo ontwikkelt zullen er ook na verschijning van dit rapport vele nieuwe toepassingsvormen ontstaan. Daarom is onderstaande inventarisatie niet uitputtend, maar geeft die primair een beeld van welk type toepassingen, binnen welk type relaties voor welk type doeleinden kunnen worden gebruikt. In deze paragraaf zal worden stilgestaan bij positieve kanten van deepfakes, in de volgende paragraaf bij de negatieve kanten. In deze paragrafen zal een bredere blik op deepfakes worden gegeven; in de hiernavolgende juridische hoofdstukken zal vervolgens met name worden ingezet op de bestudering van de inzet van deze techniek in horizontale verhoudingen.

Allereerst bieden deepfakes burgers en bedrijven de mogelijkheid om grappige filmpjes en memes te genereren of voor andere creatieve doeleinden te gebruiken. Een voorbeeld is het plaatsen van Nick Cage in andere films dan waarin hij heeft gespeeld (en dat zijn er al zo veel),[84] Nancy Pelosi ogenschijnlijk dronken een toespraak laten geven,[85] Mark Rutte en Sigrid Kaag op Temptation Island[86] of het genereren van fakeinterviews met bekende personen, het vervaardigen van een eigen journaal(item) of het afleveren van een alternatieve kersttoespraak door de Koningin.[87] Zulke deepfakes kunnen uiteraard ook op kleine schaal worden gemaakt en gebruikt, zoals kinderen die van hun ouders een grappig filmpje maken of medeburgers die van hun buurtgenoten een komische meme, foto of geluidsfragment fabriceren. Zeker als het bekendere mensen betreft, het filmpje duidelijk een deepfake is en als zodanig wordt verspreid en als de deepfake geen kwetsende of expliciete beelden of geluidsfragmenten bevat zal de video weinig schade aan anderen berokkenen. Ook zullen deepfakes worden ingezet voor (politieke) satire, door gewone burgers of cabaretiers.

Daarnaast kunnen overleden personen door deepfaketechnologie weer tot leven worden gewekt. Daarbij kan het gaan om historische figuren, zoals Salvador Dalí die plots weer tot ons spreekt[88] of de Mona Lisa die ons niet alleen uit alle hoeken aanstaart, maar ook tot ons spreekt.[89] Zowel Google[90] als Microsoft[91] hebben nieuwe methoden ontwikkeld waarbij het mogelijk is om afbeeldingen met een lage resolutie of beschadigde afbeeldingen te verbeteren zodat deze een hogere resolutie krijgen of onbeschadigd lijken. Hierdoor kunnen historische foto's verbeterd worden en veiliggesteld worden gesteld voor toekomstige generaties. Ook kunnen overleden geliefden door middel van deepfakes weer tot leven worden gewekt, bijvoorbeeld voor een speech op hun eigen begrafenis. '"It makes me so happy to see him smile again," one user said after animating a photo of her husband, who died 4 years earlier. "It's as if they are looking at you and your surroundings and seeing how much things

have changed," said another user. Reporter Joe Fitzgerald Rodriguez commented that the feature gave him a chance to see his late father's face move again after he lost the only videotape he had of him years ago. "Forget iPhones and self-driving cars," one commenter said in response to a Deep Nostalgia™ animation. "This is the moment we officially started living in the future!"'[92]

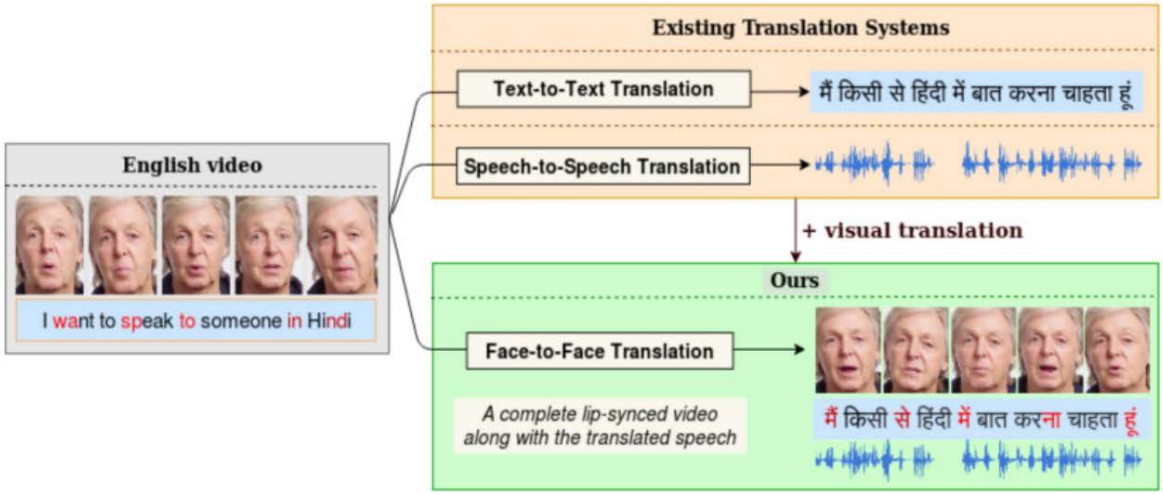
Deepfakes worden ook steeds meer ingezet in de filmbranch. In de film Furious 7, werd Paul Walker – die overleed op het moment dat hij nog bezig was met de opnames voor de film – weer tot leven werd gebracht om de film af te ronden.[93] Ook bij nog levende personen kan deepfake een uitkomst bieden, bijvoorbeeld omdat acteurs geen lange dagen op de set hoeven te blijven of voor 1 scène die niet goed is opgenomen terug moeten komen, maar hun beeltenis simpelweg kan worden ingescand en zij zodoende in allerhande scènes kunnen worden geprojecteerd. Het is daarmee in de toekomst mogelijk om acteurs geheel te vervangen; zij worden dan gevraagd om hun beeltenis en stemgeluid in te scannen, wat enkele uren duurt, waarna hun 'karakter' of 'avatar' kan worden gebruikt voor tal van producties uitgebracht door die filmstudio. Dat scheelt niet alleen geld, filmstudio's zijn niet langer gebonden aan fysieke beperkingen, zoals vrij dagen, uithoudingsvermogen of lichamelijke begrenzingen van de acteur.[94] Ook stand-ins, stuntmannen, catering en de hele entourage op een filmset kan zo verdwijnen.[95] Daarnaast zijn er nu al virtuele influencers, die niet op een daadwerkelijk persoon is gebaseerd, die grote aanhang hebben en sponsorcontracten binnenhalen.[96]

Deepfakes kunnen ook worden ingezet voor contacten met of tussen burgers, ofwel in persoonlijk ofwel in zakelijk verband. Zo kunnen bij live videobelgesprekken niet alleen personen de vorm van een kat, een hond of fictief personage aannemen,[97] als personen in een verschillende taal spreken, dan kunnen hun woorden live worden vertaald en kunnen de lippen van de persoon zo gesynchroniseerd worden dat zij gelijklopen met de vertaling. Volgens onderzoekers zal deze techniek 'significantly improve the overall user experience for consuming and interacting with multimodal content across languages'.[98] De techniek wordt onder andere ingezet voor businesscalls met China.

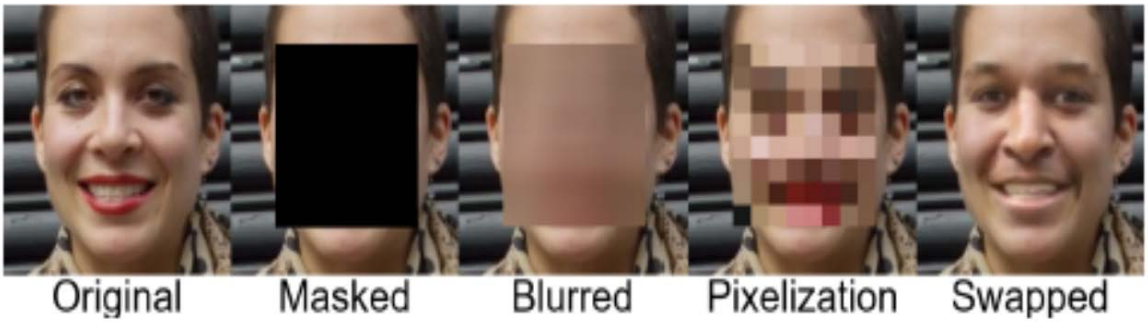
Ook kunnen deepfakes worden ingezet om de privacy van personen te waarborgen; door een andere identiteit aan te nemen kunnen zij bijvoorbeeld anoniem blijven, terwijl zij wel live aan gesprekken kunnen deelnemen voor medisch onderzoek. 'Deepfake technology could be very suitable for some medical applications. Privacy and ethical issues greatly restrict medical image and video sharing when human faces are presented. This problem is very common in neurological diseases such as movement disorders. During video recording for diagnosis, it is often inevitable to record subjects' faces. Sometimes the faces of patients are clinical manifestation of movement disorder. An alternative for sharing patient movement videos is to extract 3D keypoints and publish the keypoints data instead. However, raw videos could provide far richer information than merely keypoints. In many cases, the raw videos of patients are way more reliable and valuable than keypoints extracted by algorithms. Also, state-of-the-art keypoint detection algorithms are still far from optimal and reliable tracking. For example, Parkinson's disease is a nervous system disease

that presents movement disorders, such as bradykinesia on face and tremor. Parkinson's disease manifestation is evaluated by the Unified Parkinson's Disease Rating Scale (UPDRS). The UPDRS examinations are often videotaped for further analysis. For instance, automated keypoint detection to assess the movement without markers is an emerging field. Lack of high quality video dataset hinders the research on movement disorders. As deepfakes becomes a new option for face de-identification, video data sharing could be more applicable than before.'[99]

Binnen het medische domein bestaan meer kansen voor deepfakes.[100] Zo kunnen synthetische MRI-beelden van hersenen met tumoren worden gemaakt en kunnen vanuit die beelden algoritmen worden getraind om in de toekomst vroege vormen van kanker of hersenziektes zoals Alzheimer op te sporen. Niet alleen kan deze technologie worden gebruikt voor MRI-scans, ook is het in staat om afbeeldingen te ontwikkelen van huidlaesies die gebruikt kunnen worden door dermatologen[101] en van leverlaesies.[102] Daarnaast zouden mensen met ALS door middel van deepfake technologie de mogelijkheid kunnen behouden te spreken middels hun eigen stem.[103] Dit zou ook een oplossing kunnen zijn voor mensen met dysartrie, een spraakstoornis die wordt veroorzaakt door een beschadiging van het zenuwstelsel.[104] Ook zijn bepaalde rouwtherapieën door middel van het communiceren met een overleden echtgenoot mogelijk door deepfaketechnologie.[105]



Figuur 9: Voorbeeld real-time vervorming beeld en spraak



Figuur 10: Voorbeeld anonymisering en pseudonymisering van persoon door middel van deefake

Voor mensen met een visuele beperking is Canetroller[106] ontwikkeld, een virtuele omgeving die mensen met een visuele beperking de mogelijkheid biedt te leren navigeren door middel van stokvaardigheden. Hierbij is het mogelijk in verschillende ruimtes te navigeren, waarbij zowel binnen gebouwen kan worden genavigeerd als in de buitenwereld. Ook kunnen deepfakes mogelijkheden bieden voor mensen die aan het revalideren zijn. Zo bestaat er een 'virtual reality (VR) based cognitive neurorehabilitation system for improving the rehabilitation of stroke patients with arm and hand paresis. Using a custom, low-cost kinematic tracking system designed for clinical or home use, patients engage in task-oriented interactions with objects in a virtual environment. Our paradigm is based on the hypothesis that observed actions correlated with self-generated or intended actions activate the motor pathways by means of the so-called "mirror-system".'[107] Dergelijke deepfaketoepassingen kunnen ook helpen bij mensen met een identiteitsstoornis of slecht zelfbeeld.

Ook binnen het forensisch onderzoek bieden deepfakes mogelijkheden, bijvoorbeeld door gewelddadige incidenten op een realistische na te bootsen om zo handelingen, actoren en causale verbanden met lichamelijk letsel beter te kunnen bestuderen.[108] Ook kunnen virtuele personen worden ingezet voor het in kaart brengen van daders. Een voorbeeld was Sweetie, een virtuele avatar die in werd gezet om kinderlokkers en pederasters op de sporen. 'Sweetie is een virtueel Filipijns meisje van tien jaar. Een heel realistisch meisje dat online wordt ingezet in chatrooms en datingsites. Wanneer mannen haar benaderen voor seksueel getinte chats gaat zij met hen in gesprek. Alle informatie die dit oplevert wordt opgeslagen en gebruikt om daders te waarschuwen, op te sporen, of zelfs op te pakken en te veroordelen.'[109] In een flink aantal landen heeft de inzet van Sweetie 2.0 inderdaad tot veroordelingen geleid.[110]

Retailbedrijven kunnen virtuele modellen ontwikkelen, waarbij klanten hun eigen gezicht kunnen ruilen voor die van de virtuele modellen. Hierdoor zal het mogelijk zijn om virtueel kleding te passen, waarbij rekening wordt gehouden met het lichaamstype van de klant. De app Superpersonal[111] heeft zo'n virtuele paskamer al ontwikkeld, waarbij klanten de mogelijkheid hebben om op basis van gegevens over hun geslacht, lengte en gewicht, kleding te passen. Deze mogelijkheid wordt ook benut door online opticiens, waarbij het mogelijk is om door middel van het aanzetten van de camera op een telefoon of laptop, een bril te passen.[112] Dat deze technologie groeit, komt onder andere door de huidige pandemie. De vraag naar e-commerce is gestegen als gevolg van het wereldwijd sluiten van fysieke winkels. De verwachtingen zijn dat de omvang van de virtuele paskamermarkt de komende jaren toeneemt van ruim drie miljard (2019) naar 6,5 miljard dollar (2025).[113]

Binnen het onderwijs bieden deepfakes de mogelijkheid om interactievere lessen te geven, bijvoorbeeld door een docent geschiedenis die historische figuren tot leven kan wekken en mogelijk in de taal van de kinderen tot hen kan spreken of met hen in gesprek kan gaan.[114] Goede doelen zetten deepfakes in, bijvoorbeeld door een celebrity ogenschijnlijk in alle talen van de wereld te laten oproepen het doel te steunen,[115] door een politicus een bepaalde uitspraak te laten doen die verzet oproept[116] of door mensen te laten zien hoe een verwoeste stad er uit

ziet waar veel vluchtelingen vandaan komen door bekende steden een 'remake' te geven en burgers te laten zien hoe hun stad verwoest door oorlog eruit zou zien.[117] Tot slot kunnen politici door middel van deepfaketechnologie boodschappen verkondigen in elke gewenste taal, om binnenlandse minderheden te bereiken of in het buitenland hun boodschap beter over het voetlicht te krijgen.[118]

2.3 Negatieve toepassingen (risico's)

Ook zijn er de nodige risico's.

Een van de eerste toepassingen van deepfaketechnologie door burgers was in 2017 door een gebruiker van het platform Reddit, waarbij gezichten van bekende mensen, waaronder Taylor Swift, werden geplaatst op de lichamen van pornoactrices. Hierna explodeerde de populariteit van deze tool, waarna duizenden gebruikers creaties deelden op het platform. Vanaf februari 2018 reageerden grote platforms door toepassingen van deepfakes aan banden te leggen. Reddit verbande de deepfakes-subreddit, dit is een forum gewijd aan dit specifieke onderwerp op de website van Reddit. 'Reddit does not allow content that impersonates individuals or entities in a misleading or deceptive manner. This not only includes using a Reddit account to impersonate someone, but also encompasses things such as domains that mimic others, as well as deepfakes or other manipulated content presented to mislead, or falsely attributed to an individual or entity. While we permit satire and parody, we will always take into account the context of any particular content.'[119] Daarna volgden meerdere websites, waaronder gamingsite Discord en het afbeeldingsplatform Gfycat omdat het plaatsen van deepfakes in de vorm van porno in strijd was met het beleid/policy van de website.[120] Ook Pornhub, Twitter[121] en Facebook[122] verwijderen deepfakevideo's.

Uit onderzoek, dat in 2019 plaatsvond, bleek dat 96 procent van alle deepfakevideo's pornografisch is en gemaakt zonder wederzijds goedvinden.[123] Veruit het grootste gedeelte van de slachtoffers is vrouw en vaak ofwel een direct bekende van de dader ofwel een publiek persoon.[124] Terwijl vrouwelijke politici regelmatig slachtoffer zijn van pornografische deepfakes, gaat het bij mannen vaak om woorden die hen in de mond worden gelegd,[125] zoals Trump die België zogenaamd oproept om uit het Parijsakkoord te stappen[126] of Obama die Trump lijkt uit te maken voor een *"total and complete dipshit"*.[127] Reputatieschade kan uiteraard ook worden toegebracht aan merken en bedrijfsnamen, zowel van grote multinationals als op lokaal niveau, bijvoorbeeld als twee concurrerende eenmanszaken elkaar zwart proberen te maken.

In april 2020 doken deepfake-audioclips op waarin een text-to-speech-model werd gebruikt dat was getraind in de spraakpatronen van Jay-Z om zich zo voor te doen als Jay-Z die de monoloog *'To be, or not to be'* uit Shakespeare's Hamlet citeerde.[128] Roc Nation, Jay-Z's label, deed vervolgens een verzoek tot verwijdering wegens het schenden van auteursrechten. Youtube reageerde aanvankelijk door de video te verwijderen, maar later is de video opnieuw online gekomen omdat de eiser onvoldoende gronden had aangevoerd om aan te tonen dat het materiaal inderdaad onrechtmatig was. Het Amerikaanse Intellectueel Eigendomsrecht is volgens de maker van de video niet eenduidig op dit punt.[129]

Meer in het algemeen is de vraag in hoeverre portret- of persoonlijkheidsrechten rusten op beelden van overleden personen en in hoeverre nazaten daarvan daar een beroep op kunnen doen. Niet iedereen vindt het een fijne gedachte om jaren na zijn dood nog tot leven te worden gewekt, terwijl zeker voor bekende mensen, dit de nabestaande wel een goed verdienmodel kan opleveren. Dit raakt aan de discussie over post *mortem privacy*; in hoeverre kunnen overleden personen een beroep doen op reputatie- en persoonlijkheidsrechten?[130] 'What does it mean to resurrect the dead, and not only bring them back to life to communicate with them, but to speak *through* them? Digital communications are often meant to be short-term, real-time, and immediate, but through what I think of as "platform temporality," they can be collected and preserved, and potentially passed on to loved ones and future descendants. There is also some amount of anxiety attached to the capacity for data to live on past people's physiological selves; it is hard to control data during life and it is nearly impossible to do so after one's death. Living on through social media accounts is one thing, but actually using A.I. to replicate a person's personality in perpetuity is something different entirely, especially when it comes to deepfakes.'[131]

Deepfakes kunnen ook worden ingezet voor financieel gewin, zoals het manipuleren van markten. Nadat in 2019 een vals berichtje de ronde deed op de communicatie-app Whatsapp, waarin stond dat Metro Bank niet meer liquide was, gingen mensen massaal naar de Metro Bank om al hun geld en sieraden op te eisen. Dit leidde er uiteindelijk toe dat het aandeel met 9% daalde.[132] Criminelen kunnen zich met gebruikmaking van een deepfake voordoen als de CEO van een beursgenoteerd bedrijf, waarbij de CEO bijvoorbeeld voor dat bedrijf schadelijke uitspraken doet waardoor de aandelenkoers daalt. Europol noemt deepfakes dan ook een aanzienlijk gevaar voor onder meer 'perpetrating extortion and fraud, facilitating document fraud, falsifying online identities and fooling KYC [Know Your Customer] mechanisms, falsifying or manipulating electronic evidence for criminal justice investigations, disrupting financial markets' en bijvoorbeeld het ontfutselen van bedrijfsgeheimen door middel van deepfakes.[133] Een voorbeeld van fraude komt uit 2019, toen voor het eerst, althans voor zover bekend, door middel van een deepfake audio-fragment iemand werd opgelicht. Een CEO van een energiebedrijf in het Verenigd Koninkrijk maakte €220.000 over naar een Hongaars bankrekening, omdat hij in de veronderstelling was dat hij aan de telefoon zat met zijn baas, de baas van het Duitse moederbedrijf, die hem zulks leek op te dragen.[134]

Middels een deepfakevideo zou een politicus – ten tijde van verkiezingen – zijn tegenstander in diskrediet kunnen brengen of een politiek schandaal aan kunnen wakkeren. Zo leidde een deepfakevideo van voormalige premier van Italië, Matteo Renzi, waarin hij een mede-politici beledigde, tot publieke verontwaardiging.[135] 'The deepfake video refers to Renzi's decision Sept. 17 to leave the Democratic Party and form his own party. In the parody, the supposed Renzi is seen talking when he thinks he is off air. He discusses the reaction of various politicians, including Prime Minister Giuseppe Conte; Luigi Di Maio, leader of the Five Star Movement; and Italy's president, Sergio Mattarella. The video is so outrageous that it is clearly a parody, but deepfake technology makes it look incredibly realistic. So when people started sharing it online, claiming

that it was a real video, quite a few social media users fell for it and were outraged by what they saw as Renzi's bad behavior.'[136]

Zulke lastercampagnes zouden als gevolg kunnen hebben dat burgers hun stem bij verkiezingen uitbrengen mede op basis van desinformatie.[137] Zo wijst de Minister van Binnenlandse Zaken en Koninkrijksrelaties in een Kamerbrief op het gevaar van desinformatie voor de Nederlandse democratische rechtsorde.[138] 'Zoals ook in mijn eerdere brieven genoemd is (heimelijke) politieke beïnvloeding geen nieuw fenomeen. Het omvat de integrale, veelal heimelijke inzet van (drog-) argumenten, selectieve informatie en desinformatie (omtrent politiek gevoelige thema's) ten behoeve van het realiseren van politieke doeleinden richting een vooraf bepaald publiek. De opkomst van het internet heeft het echter wel een nieuwe dynamiek gegeven: brede verspreiding van desinformatie en nepnieuws kan makkelijk, anoniem, snel en goedkoop. Zoals aangegeven in de jaarverslagen van de AIVD zijn er statelijke actoren die zich richten op Nederland en onder meer de intentie en capaciteit hebben om zich te mengen in democratische processen. Zij zijn geïnteresseerd in politieke besluitvorming en hoogwaardige technologische kennis (sectoren ICT, maritieme technologie, biotechnologie en lucht- en ruimtevaart). Daarbij wordt een palet aan middelen ingezet, zoals de inzet van klassieke inlichtingenofficieren, omvangrijke en hardnekkige digitale aanvallen op publieke en private organisaties en de verspreiding van desinformatie.'[139]

Deepfakes kunnen ook worden gebruikt om binnenlandse of interstatelijke conflicten uit te lokken. In 2018 was de president van Gabon, Ali Bongo, voor medische behandelingen het land uit. Mensen werden achterdochtig over het welzijn van de president, waarbij sommigen vermoedden dat hij reeds was overleden. Na een aantal maanden kondigde de vicepresident aan dat de president een beroerte had gehad. Maatschappelijke groeperingen en het publiek bleven zich afvragen waarom de president nog niet in het openbaar was verschenen, waarna de regering een video van de president uitbracht. Velen geloofden niet dat de video echt was en een week nadat de video was gepubliceerd, ondernam het leger van Gabon een staatsgreep, die uiteindelijk mislukte.[140] De relatie tussen verschillende staten kan ook onder spanning komen te staan door een deepfake. In 2020 eiste de Australische premier, Scott Morrison, bijvoorbeeld excuses van China, nadat een woordvoerder van het Chinese ministerie van Buitenlandse Zaken een nepfoto op Twitter had geplaatst waarop een Australische soldaat te zien was die een mes tegen de keel van een Afghaans kind hield, wat China echter weigerde.[141] De vrees is dat bijvoorbeeld Russische trollen op termijn democratische verkiezingen in andere landen zullen proberen te beïnvloeden door het verspreiden van fake video- en audioberichten.

Europol vreest daarnaast dat deepfakes zullen worden gebruikt voor 'distributing disinformation and manipulating public opinion, inciting acts of violence toward minority groups, supporting the narratives of extremist or even terrorist groups, and, stoking social unrest and political polarization'.[142] Desinformatie verspreiden kan niet alleen effecten hebben voor concrete personen of groepen, maar op den duur ook leiden tot wantrouwen in media. Daarmee kan de trend dat mensen nog meer in hun eigen waarheid gaan geloven worden versterkt, wat op zijn beurt weer een effect kan hebben op de

maatschappij en de democratische rechtsstaat, omdat groepen steeds meer langs en tegen elkaar leven.[143]

Als laatste voorbeeld van hoe deepfakes op termijn grotere, meer systemische vragen kan oproepen kan worden verwezen naar het gebruik van bewijsmateriaal in de rechtszaak. In 2020 legde de moeder in een Britse voogdijzaak een deepfake-audiobestand voor aan de rechter. In deze opname was te horen hoe de vader de moeder bedreigde, de moeder had dit bestand zelf gefabriceerd om haar beweringen – dat de vader gewelddadig was – te ondersteunen. Het dossier is forensisch onderzocht, waarna werd bewezen dat het audiofragment vals was.[144] In dit geval is uitgekomen dat er sprake is van een deepfake, maar doordat de kwaliteit van deepfakes steeds beter wordt en minder goed te detecteren is, is de vraag hoe dit in de toekomst zal zijn. Als geen scherpe controle op bewijsmateriaal plaatsvindt kan dit ten gevolg hebben dat mensen ten onrechte worden veroordeeld voor delicten die zij niet hebben begaan of dat hen de voogdij over hun kinderen wordt ontzegd op basis van onechte beeld- of geluidsfragmenten. Als echter voortaan al het bewijsmateriaal op echtheid moet worden gecontroleerd kan die een enorme toename in kosten en tijdsduur voor rechtszaken met zich brengen.

2.4 Maatschappelijke effecten

Tot slot zijn er algemene ontwikkelingen die de specifieke toepassingen van deepfakes en de concrete gevolgen daarvan in individuele gevallen overstijgen. Het is onmogelijk om op deze gevolgen de precieze vinger te leggen en een directe causale relatie aan te tonen met deepfakes, onder meer omdat deze maatschappelijke gevolgen ook samenhangen met en afhankelijk zijn van tal van andere technische, maatschappelijke en politieke ontwikkelingen. Daarom zal kort worden geschetst aan welke mogelijke individu-overstijgende risico's deepfakes bijdragen. Deze hangen veelal samen met maatschappelijke ontwikkelingen die reeds gaande zijn.

Er is reeds een maatschappelijke tendens, offline en zeker online, om disrespectvolle uitingen te doen aan het adres van vrouwen en meisjes.[145] Niet zelden hebben publieke uitingen over vrouwen op Twitter of andere sociale media een negatieve of misogyne ondertoon. Daarbij komt dat het vrouwenlichaam, meer dan het mannenlichaam, wordt geobjectiveerd en als lustobject wordt gezien. Daaraan zijn ook tal van al dan niet realistische schoonheidsidealen gekoppeld. Tot slot geldt er nog steeds een andere seksuele moraal voor vrouwen dan voor mannen. Tieners en jongvolwassen mannen maken bangalijstjes van vrouwelijke leeftijdgenoten, terwijl slutshaming en het versturen van naaktbeelden zeer ingrijpende gevolgen kan hebben voor met name jongvolwassen meisjes. Die gevolgen zijn niet minder als het gaat om beelden die foutief aan een vrouw worden gekoppeld, ook niet als reeds lange tijd duidelijk is dat het om een onjuiste toeschrijving gaat. De video of foto in kwestie kan iemand nog jarenlang achtervolgen en de negatieve sociale en maatschappelijke gevolgen ook. In extremo leiden zulke incidenten tot zelfmoord.[146]

Het is niet moeilijk voor te stellen hoe deepfakes deze maatschappelijke tendens kan versterken en vergroten. Bij veel van de incidenten die nu plaatsgrijpen gaat het om afbeeldingen of video's gemaakt in de privésfeer die door een

van beide partners openbaar wordt gemaakt, vaak bij wijze van wraak. Door de introductie van deepfaketechnologie is het niet langer nodig om dergelijke beelden ook echt te maken; ze kunnen door nu al op de consumentenmarkt beschikbare deepfake-apps worden genereerd. Deze apps geven op basis van een afbeelding van een persoon waarop die met kleding te zien is een inschatting van hoe iemand er naakt uit ziet. Dergelijke naaktbeelden, al dan niet gebruikt in een video waarin een persoon seksuele handelingen lijkt te verrichten, kunnen nog explicieter van aard zijn en niet van echt te onderscheiden. Ze kunnen worden vervaardigd door elke persoon met toegang tot een foto van een vrouw of meisje, zoals klasgenoten, buurtgenoten of collega's. Dit kan de sociale onveiligheid voor vrouwen in het algemeen vergroten en opgroeiende meisjes in het bijzonder vergroten. Ook kunnen dergelijke seksueel getinte deepfakes bijdragen aan de objectivering van het vrouwenlichaam en aan onrealistische schoonheidsidealen.

Een andere tendens die gaande is betreft de vervaging tussen wat echt is en wat nep. Dit wordt beschreven als dat de huidige wereld zich in een post-truth tijdperk bevindt,[147] waarin leugens en waarheden steeds moeilijker van elkaar te onderscheiden zijn[148] en waarin meningen en objectieve feiten met elkaar om de eerste plaats strijden.[149] Dit sluit aan bij de verkokering en sociale stratificatie van de samenleving. Onder meer door de zogenoemde echo-chambers of filter bubbles[150] op het internet, waarin mensen meningen, nieuws, advertenties en zoekresultaten te zien krijgen die aansluiten bij hun bestaande opvattingen en wereldbeeld. Hierdoor is er een tendens gaande waarin groepen steeds meer in hun eigen waarheid gaan geloven en er minder ruimte is voor nuance. Dit leidt tot polarisatie. Als voorbeeld wordt vaak verwezen naar de Verenigde Staten waar links en met name rechts in een steeds eigenzinnigere versie van de waarheid geloofd. Zoals benoemd in de inleiding bij dit rapport gelooft nu al een substantieel deel van de Amerikaanse bevolking in evidente leugens van Trump, nog zonder dat daar realistische deepfakes aan te pas zijn gekomen.[151]

Ook hierbij is eenvoudig voor te stellen hoe deepfakes deze tendens zullen vergroten. Deepfakes zijn eenvoudig te vervaardigen en te verspreiden onder bepaalde doelgroepen. Het is bekend dat als de boodschap of inhoud van een bepaald bericht aansluit bij het toch al bestaande wereldbeeld van de consument, deze eerder voor waar zal worden aangenomen dan als die daar strijdig mee is. Door de te verwachten aanzienlijke toename in het aantal nepberichten en deepfakes zal de waarheid en authentieke berichtgeving op termijn ondergesneeuwd kunnen raken, zeker onder groepen die moeite hebben met berichtgeving van de zogenoemde Mainstream Media. Doordat de ene groep in authentieke berichtgeving gelooft en de andere groep in daadwerkelijke fake-berichten, of twee of meerdere groepen in verschillende fake-versies van de werkelijkheid, sluiten hun wereldbeelden elkaar steeds meer uit.

Deze verspreiding van misinformatie heeft gevolgen voor drie belangrijke maatschappelijke instituten:

Ten eerste de democratie. Voor dit gevaar is momenteel reeds de meeste aandacht. Hierbij wordt met name gewezen op incidenten waarbij buitenlandse mogendheden, en dan met name Rusland, fakeberichten en trollen lijken in te zetten

om democratische processen te beïnvloeden. Een aantal landen en meerdere staten van de Verenigde Staten, hebben reeds wetgeving aangenomen om zich hiertegen te wapenen. Ook is duidelijk dat belangengroepering binnen de landsgrenzen deepfakes inzetten om hun politieke wensen over het voetlicht te brengen of door hun gesteunde kandidaten in een goed daglicht te stellen. Ook wijzen experts, onder meer in de voor deze studies gehouden interviews, erop dat staten ook concrete, voor hen relevante, besluitvorming in andere landen trachten te beïnvloeden door de verspreiding van nepnieuws. Door deze toepassingen kunnen deepfakes een potentieel zeer ontwrichtende werking hebben op het democratisch proces als zodanig.

Ten tweede de pers. Nu al kost het de pers moeite om alle berichten goed op authenticiteit te controleren en worden daar significante kosten en manuren aan besteed. Een klein maar veelzeggend voorbeeld is de voetballer die tijdens een EK-wedstrijd van het veld werd gedragen vanwege een medische conditie, waarna al vrij snel een foto op Twitter verscheen waaruit zou blijken dat hij zich in ieder geval nog in levende conditie bevond. Toch duurde het een flinke tijd voordat officiële media-kanalen gewag maakten van deze foto, omdat er ter degen rekening mee werd gehouden dat het hier om een gefabriceerd beeld ging.[152] In een wereld waarin iedere burger toegang heeft tot deepfaketechnologie en binnen enkele seconden een nep-video, -foto of -audiobestand op het internet kan verspreiden is de vraag hoe media praktisch gezien ervoor kunnen zorgen dat de door hun overgenomen berichten kloppend zijn. Kwaliteitsmedia die in dergelijke procedures investeren lopen niet alleen het risico dat zij minder winst maken vanwege de kosten van dergelijke procedures, maar ook het risico 'achterhaald' te raken, omdat andere media, die minder zorgvuldigheidseisen hanteren, altijd sneller zullen zijn met berichtgeving, die dan ook nog eens sensationeler van aard is.

Ten derde de rechtstaat. Deze kan onder druk komen te staan vanwege deepfakes vanwege een aantal redenen. (1) Processen kunnen langer duren, omdat partijen altijd kunnen beweren dat tegen hen geleverd bewijs nep en gefabriceerd is. Dit werpt een verdere drempel op voor het voeren van juridische processen, die toch al significant is. (2) Het gevaar is dat de rechter content onterecht voor waar zal aannemen en dit tot een onterechte veroordeling leidt. (3) Een veroordeelde kan, na een rechtelijke uitspraak, altijd publiekelijk zijn onschuld volhouden door te beweren dat de rechter in een fake-bericht is gestonken. (4) Bij bepaalde delicten is de suggestie al genoeg voor publieke verontwaardiging, zoals reeds is gebleken bij een celebrity die in een fake-video een dier leek te mishandelen en doodsbedreigingen moest dulden. Dergelijke verontwaardiging is ook voorstelbaar bij bepaalde seksuele fake-video's, bijvoorbeeld als het gaat om dieren of kinderen, of extremere seksuele handelingen. Als het OM dan vervolgens niet tot vervolging overgaat, omdat het tot het oordeel is gekomen dat de video nep is, is het niet ondenkbaar dat deze beslissing minder aandacht zal krijgen dan het oorspronkelijke nepbericht, zodat de waarheid de leugen niet altijd zal achterhalen.

2.5 Typologie

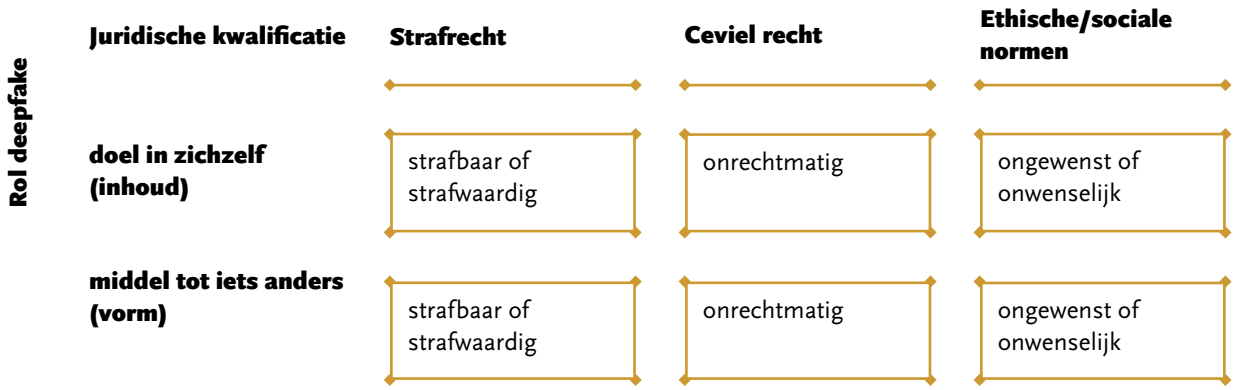
Uit het voorgaande is de volgende categorisering te maken van toepassingen en gevolgen van deepfaketechnologie.[153]

	Positieve toepassingen	Negatieve toepassingen
Creëren van nieuw persoon	Gebruik van fake-persoon voor infiltratie; gebruik fake-kinderporno om pedoseksuelen te behandelen; avatars voor in spellen;	Kinderporno; fake nieuws;
Bestaande personen nieuwe handelingen laten verrichten/ woorden in de mond leggen/ etc.	Beeldbellen in andere taal; satire t.a.v. politici en bekende personen; acteurs gevaarlijke stunts laten verrichten; politici die minderheden in hun taal toespreken; nabootsen plaats delict; medische toepassing voor mensen met een verstoord zelfbeeld; klanten virtueel kleding laten dragen;	Porno zonder toestemming; kinderporno; fraude; misleiding in de rechtszaal; verspreiden fake nieuws; politieke incidenten veroorzaken; iemand in diskrediet brengen; iemands beeld of gelijkenis exploiteren; aanzetten tot haat of geweld tegen personen of groepen;
Dode personen weer tot leven wekken	Communiceren met overleden partner; zien hoe overgrootouders er uit gezien zouden hebben; overleden persoon laten figureren op eigen begrafenis; overleden artiest/acteur laten figureren, b.v. in film of als hologram in concert; overleden persoon college/rondleiding laten geven	Porno; kinderporno; iemands beeltenis commercieel exploiteren;
Het vermengen van beelden/geluiden/etc. van diverse bestaande personen	Satire; pseudonimisering van personen;	Porno; kinderporno; creëren onrealistisch zelfbeeld/schoonheidsideaal
Anonimiseren van personen	Anonieme getuige middels Deepfake laten verschijnen in rechtszaal; anonieme patiënt laten meedoen in een trial	

Figuur 11: Categorisering deefakes aan de hand van gevolgen

Om verder vorm te geven aan dit onderzoek dient te worden bepaald welke vormen van deepfakes en de verspreiding relevant zijn voor een nadere juridische analyse en welke worden uitgesloten (bijvoorbeeld omdat de toepassing onschadelijk is en niet in strijd met de AVG of het WvSr). Deze dienen dan als input voor de juridische analyse. Daarbij zullen de verschillende typen deepfakes op het gehele juridische continuüm worden beschreven. Een Deepfake kan in een van de relevante rechtsgebieden vallen, maar ook in meerdere of allemaal. Juist de deepfakes die zich op het grensvlak van verschillende juridische domeinen afspelen zijn relevent en roepen complexe vraagstukken op omtrent de interactie tussen die rechtsgebieden. Een vertrekpunt voor het juridisch kwalificeren van deepfakes zal zijn:

- 1. Deepfakes als intrinsiek strafwaardige content;
- 2. Deepfakes als vehikel om een strafwaardig feit te plegen;
- 3. Deepfakes als intrinsiek onrechtmatige content;
- 4. Deepfakes als vehikel voor onrechtmatige content;
- 5. Deepfakes als ongewenste (maar geen strafbare/onrechtmatige) content;
- 6. Deepfakes als legitieme content.



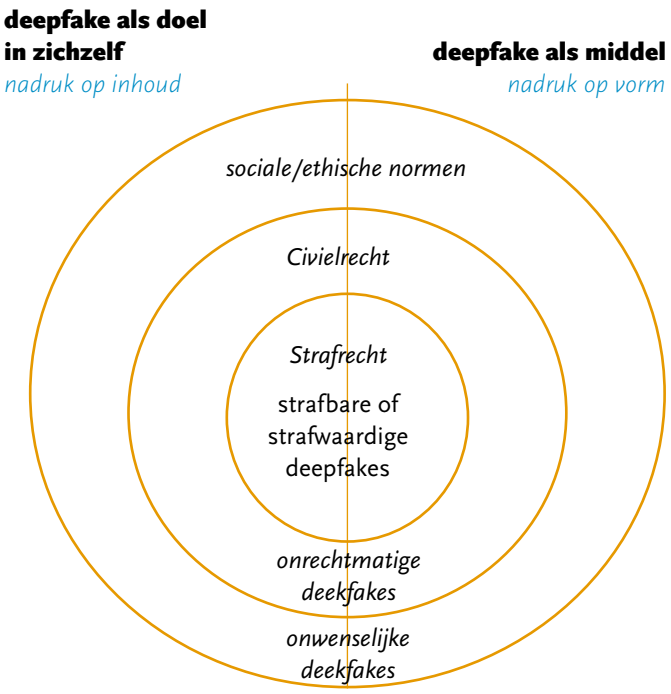
Figuur 12: Juridische kwalificatie deepfakes

In het volgende hoofdstuk zal nader worden bekeken in hoeverre het huidige juridische kader al afdoende regels stelt om deze risico's te ondervangen. Daartoe is de volgende schematisering en onderverdeling van belang.

Dit kan als volgt gevisualiseerd worden (figuur 13), waarbij duidelijk wordt dat wat strafbaar of strafwaardig is, doorgaans ook onrechtmatig is, en wat strafbaar of strafwaardig of onrechtmatig is, doorgaans ook sociale of ethische normen overschrijdt. Dat geldt andersom niet. Een schending van een sociaal of ethische norm is niet altijd onrechtmatig en een onrechtmatige handeling is niet altijd strafbaar of strafwaardig. Daarbij geldt uiteraard dat wat als strafbaar of strafwaardig dient te worden gezien vaak veel eenduidiger is vast te stellen dan of iets tegen sociale of ethische normen ingaat, wat afhangt van persoonlijke of culturele opvattingen en normatieve uitgangspunten. Dit onderscheid is belangrijk omdat het Openbaar Ministerie slechts verantwoordelijk is voor de handhaving van strafbare feiten, de Autoriteit Persoonsgegevens zich zowel kan uitlaten over strafbare als over onrechtmatige gegevensverwerking en dat internet intermediairs niet alleen de plicht kunnen hebben om strafbare of onrechtmatige feiten van hun platform te weren, maar ook een maatschappelijke plicht hebben om content die sociale of ethische normen overschrijdt te verwijderen.

Bij deze categorisering wordt een initieel onderscheid gemaakt tussen deepfakes als een doel op zich (het zaaien van verwarring/angst, het kwetsen van mensen, etc.) en deepfakes die primair instrumenteel zijn aan andere feiten die ofwel stafwaardige, ofwel onrechtmatige of onwenselijke handelingen betreffen (bijvoorbeeld deepfakes om via identiteitsdiefstal fraude te plegen, deepfakes om haat te zaaien, etc.). Bij de laatste toepassingsvorm is het onderliggende feit vaak reeds genormeerd (in die zin is deepfake niet anders dan andere 'listige kunstgrepen' die historisch reeds bekend zijn, zoals het verkleden of fotoshoppen). Met name bij die eerste categorie komen nieuwe juridische vraagstukken in materieelrechtelijke zin naar voren, alhoewel ook hierbij uiteraard geldt dat de twee categorieën niet absoluut van elkaar te onderscheiden zijn.

In de volgende hoofdstukken zal met name worden ingezoomd op de deepfakes met risico's en gevaren, omdat de besproken rechtsgebieden ten doel hebben om deze risico's te vermijden of te mitigeren. Daarbij zal met name aandacht worden geschonken aan deepfakes in horizontale verhoudingen en zal worden bekeken in hoeverre



Figuur 13: Categorisering van onwenselijke vormen van deepfakes

het huidige juridische kader afdoende waarborgen biedt. Daarbij geldt dat veel van de voorbeelden die in paragraaf 2.3 aan bod kwamen vanuit meerdere rechtsgebieden bezien kunnen worden. Fakepornovideo's kunnen vanuit het strafrecht, het gegevensbeschermingsrecht, het recht op eer en goede naam, het onrechtmatige-daadsrecht en zelfs het portretrecht worden bekeken. Identiteitsvervalsing ten behoeve van oplichting kan ook in strijd zijn met de rechtsregels uit ieder van deze juridische domeinen. Meineed plegen door valse audiofragmenten van de tegenpartij in te brengen evenzo. Aanzetten tot haat tegen bijvoorbeeld minderheden is een strafbaar delict, maar ook is het door het Europees Hof voor de Rechten van de Mens onder het recht op eer en goed naam geschaard, worden er persoonsgegevens voor verwerkt en kan immateriële of materiele schade die uit die aanzet volgt onder het Burgerlijk Wetboek worden verhaald. Vaak zullen dus meerder rechtsgebieden van toepassing zijn.

2.6 Conclusie

Dit hoofdstuk heeft een overzicht gegeven van de techniek achter deepfakes, de verschillende toepassingsmogelijkheden en de gevolgen daarvan. 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. Negatieve toepassingen betreffen onder meer 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 deze concrete gevolgen van concrete inzet kunnen deepfakes ook belangrijke maatschappelijke gevolgen hebben. Daarbij valt te denken aan een afnemend vertrouwen in de media, de politiek en de rechtsspraak, polarisatie en belemmeringen voor het functioneren van deze instituten door de hoeveelheid nepmateriaal dat wordt gecreëerd en verspreid. Ook hebben de deepfake pornotoepassingen niet zelden een negatief effect op vrouwen en hun maatschappelijke positie.

Voetnoten hoofstuk 2

- 46 <https://nos.nl/artikel/2370961-deepfake-achtige-app-brengt-oude-portretten-echt-bizar-tot-leven.html>.
- 47 <https://beebom.com/best-deepfake-apps-websites/>.
- 48 <https://nos.nl/nieuwsuur/video/2371746-we-zijn-op-een-punt-dat-je-niet-meer-zeker-weet-wat-echt-is-en-wat-nep.html>.
- 49 Maras, M. H., & Alexandrou, A. (2019). Determining authenticity of video evidence in the age of artificial intelligence and in the wake of deepfake videos. *The International Journal of Evidence & Proof, 23(3), 255-262, par. 1. Hurst E. (2019). How can the law deal with Deepfake?*. Allaboutlaw. <https://www.allaboutlaw.co.uk/commercial-awareness/legal-spotlight/how-can-the-law-deal-with-deepfake->.
- 50 Mann A. (2019). *Deepfake AI: Our Dystopian Present.* Livescience. https://www.livescience.com/deepfake-ai.html
- 51 Durall, R., Keuper, M., Pfreundt, F. J., & Keuper, J. (2019). Unmasking deepfakes with simple features. arXiv preprint arXiv:1911.00686, par 2.
- 52 Onderstaande afbeeldingen afkomstig van: <https://github.com/iperov/DeepFaceLab/issues/892>. <https://mrdeepfakes.com/forums/thread-legacy-guide-deep-facelab-1-0-guide>. <https://www.wired.com/story/deepfakes-getting-better-theyre-easy-spot/>. <https://www.storypick.com/mr-bean-and-trump-funny-deepfake-video/>.
- 53 Goodfellow, I. et al. (2014). Generative adversarial networks. *arXiv preprint arXiv:1406.2661.*
- 54 Brock, A., Donahue, J., & Simonyan, K. (2018). Large scale GAN training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096,* p. 2.
- 55 Goyal S. (2019). GANs — A Brief Introduction to Generative Adversarial Networks. Medium. <https://medium.com/analytics-vidhya/gans-a-brief-introduction-to-generative-adversarial-networks-f06216c7200e>.
- 56 Feng, J. et al. (2020). Generative adversarial networks based on collaborative learning and attention mechanism for hyperspectral image classification. *Remote Sensing, 12*(7), 1149.
- 57 https://meraju.com/fakelab-a-deepfake-audio-detection-tool/
- 58 Goyal S. (2019). GANs — A Brief Introduction to Generative Adversarial Networks. *Medium*. <https://medium.com/analytics-vidhya/gans-a-brief-introduction-to-generative-adversarial-networks-f06216c7200e>.
- 59 McDonald G. (2018). Seeing Isn't Believing: This New AI System Can Create "Deep Fake" Videos. *Seeker.* <https://www.seeker.com/artificial-intelligence/this-new-ai-system-can-create-convincing-deep-fake-videos>.
- 60 Bansal, A., Ma, S., Ramanan, D., & Sheikh, Y. (2018). Recycle-gan: Unsupervised video retargeting. In Proceedings of the European conference on computer vision (ECCV) (pp. 119-135), introduction.
- 61 <https://www.youtube.com/watch?v=ehD3C60i6l-w&amp&feature=youtu.be>.
- 62 Hurst E. (2019). How can the law deal with Deepfake?. *Allaboutlaw.* <https://www.allaboutlaw.co.uk/commercial-awareness/legal-spotlight/how-can-the-law-deal-with-deepfake->.
- 63 Zakharov, E., Shysheya, A., Burkov, E., & Lempitsky, V. (2019). Few-shot adversarial learning of realistic neural talking head models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 9459-9468), par 5.*
- 64 Chan, C., Ginosar, S., Zhou, T., & Efros, A. A. (2019). Everybody dance now. In Proceedings of the IEEE/CVF *International Conference on Computer Vision* (pp. 5933-5942).
- 65 Burt C. (2019). *DataGrid develops AI to generate whole-body images of nonexistent people. Biometricupdate.com.* <https://www.biometricupdate.com/201905/datagrid-develops-ai-to-generate-whole-body-images-of-nonexistent-people>.
- 66 Durall, R., Keuper, M., Pfreundt, F. J., & Keuper, J. (2019). Unmasking deepfakes with simple features. arXiv preprint *arXiv:1911.00686*. Abstract.
- 67 Y. Li, M.-C. Chang, and S. Lyu. In ictu oculi: Exposing ai generated fake face videos by detecting eye blinking. arXiv preprint *arXiv:1806.02877,* 2018.
- 68 <https://www.forensischinstituut.nl/forensisch-onderzoek/prnu-compare-professional>.

- 69 Marra, F., Gragnaniello, D., Verdoliva, L., & Poggi, G. (2019, March). Do gans leave artificial fingerprints?. In 2019 IEEE Conference on Multimedia Information Processing and *Retrieval* (MIPR) (pp. 506-511). IEEE, Introduction.
- 70 Cozzolino, D., & Verdoliva, L. (2018). Noiseprint: a CNN-based camera model fingerprint. arXiv preprint *arXiv:1808.08396,* par. 2.
- 71 <https://www.albany.edu/news/92306.php>.
- 72 Brock, A., Donahue, J., & Simonyan, K. (2018). Large scale GAN training for high fidelity natural image synthesis. arXiv preprint *arXiv:1809.11096*. Waarbij wordt verwezen naar Amerini, I., and Caldelli, R. (2020, June). Exploiting prediction error inconsistencies through LSTM-based classifiers to detect deepfake videos. In *Proceedings of the 2020 ACM Workshop on Information Hiding and Multimedia Security* (pp. 97-102).
- 73 Korshunov, P., & Marcel, S. (2019). Vulnerability assessment and de- tection of deepfake videos. In *The 12th IAPR International Conference on Biometrics* (ICB), pp. 1-6.
- 74 X. Yang, Y. Li, and S. Lyu. Exposing deep fakes using inconsistent head poses. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 8261–8265. IEEE, 2019.
- 75 Durall, R., Keuper, M., Pfreundt, F. J., & Keuper, J. (2019). Unmasking deepfakes with simple features. *arXiv preprint arXiv:1911.00686,* par. 2.
- 76 Nguyen, T. T., Nguyen, C. M., Nguyen, D. T., Nguyen, D. T., & Nahavandi, S. (2019). Deep Learning for Deepfakes Creation and Detection: A Survey. arXiv preprint arXiv:1909.11573, par 3.
- 77 Schroepfer M. (2019). Creating a dataset and a challenge for deepfakes. *Ai.Facebook*. <https://ai.facebook.com/blog/deepfake-detection-challenge/>.
- 78 Canton Ferrer C. et al. (2020). Deepfake Detection Challenge Results: An open initiative to advance AI. *Ai.Facebook*. <https://ai.facebook.com/blog/deepfake-detection-challenge-results-an-open-initiative-to-advance-ai/>.
- 79 Burt T. (2020). New Steps to Combat Disinformation. *Microsoft*. <https://blogs.microsoft.com/on-the-issues/2020/09/01/disinformation-deepfakes-news-guard-video-authenticator/>.
- 80 Rossler, A. et al. (2019). Faceforensics++: Learning to detect manipulated facial images. *In Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 1-11), figure 1. Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., & Ferrer, C. C. (2020). The deepfake detection challenge dataset. arXiv preprint arXiv:2006.07397.
- 81 SentinelOne (2019). What is a Hash? (And How Does It Work?). *SentinelOne*. <https://www.sentinelone.com/blog/what-is-hash-how-does-it-work/>.
- 82 Ozdemir D. (2021). Teenager's AI project for detecting Deepfak videos wins Award. *Interesting Engineering*. <https://interestingengineering.com/teenagers-ai-system-for-detecting-deepfake-videos-wins-award>.
- 83 Vaccari, C., & Chadwick, A. (2020). Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media+ Society, 6*(1).
- 84 <https://www.youtube.com/watch?v=BU9YAHigNx8>.
- 85 <https://www.youtube.com/watch?v=sDOo5nDJwgA>. Dit filmpje is overigens niet noodzakelijk een Deepfake – het is gemaakt door het geluid en het beeld wat langzamer af te spelen en het stemgeluid, dat als het vertraagd wordt afgespeeld lager van toon wordt, weer op natuurlijke toonhoogte te krijgen. Wel geeft het een indruk van hoe invloedrijk een filmpje kan zijn, dan niet eens met geavanceerde middelen is gegenereerd.
- 86 <https://www.youtube.com/watch?v=noCwnzO7Agg>.
- 87 <https://www.youtube.com/watch?v=IvY-Abd2FfM>.
- 88 Lee D. (2019). Deepfake Salavador Dalí takes selfies with museum visitors. *The Verge*. <https://www.theverge.com/2019/5/10/18540953/salvador-dali-lives-deepfake-museum>.
- 89 <https://www.youtube.com/watch?v=P2uZF-5F1wI>. BBC (2019). Mona Lisa 'brought to life' with deepfake AI. BBC News. <https://www.bbc.com/news/technology-48395521>.
- 90 Berthelot, D., Milanfar, P., & Goodfellow, I. (2020).

Creating high resolution images with a latent adversarial generator. arXiv preprint *arXiv:2003.02365*.

- 91 Wan, Z. et al. (2020). Bringing old photos back to life. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 2747-2757).
- 92 <https://blog.myheritage.com/2021/02/deep-nostalgia-goes-viral/>.
- 93 <https://www.hollywoodreporter.com/behind-screen/how-furious-7-brought-late-845763>.
- 94 In The Irishman werd door middel van digitale techniek en het gezicht van Robert de Niro 30 jaar jonger gemaakt.
- 95 Over deze ontwikkeling werd reeds in 2013 een film gemaakt <https://www.imdb.com/title/tt1821641/>.
- 96 <https://influencermatchmaker.co.uk/blog/virtual-influencers-what-are-they-how-do-they-work>.
- 97 <https://www.youtube.com/watch?v=qcnnI6HD6DU>.
- 98 KR, P., et al. (2019, October). Towards automatic face-to-face translation. In *Proceedings of the 27th ACM International Conference on Multimedia* (pp. 1428-1436).
- 99 Zhu, B., Fang, H., Sui, Y., & Li, L. (2020, February). Deepfakes for Medical Video De-Identification: Privacy Protection and Diagnostic Information Preservation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 414-420).
- 100 Snow J. (2018). Deepfakes for good: Why researchers are using AI to fake health data. *Fast Company*. <https://www.fastcompany.com/90240746/deepfakes-for-good-why-researchers-are-using-ai-for-synthetic-health-data>.
- 101 Baur, C., Albarqouni, S., & Navab, N. (2018). Generating highly realistic images of skin lesions with GANs. In OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis (pp. 260-267). Springer, Cham.
- 102 Frid-Adar, M., et al. (2018). GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. *Neurocomputing*, 321, 321-331.
- 103 <https://teamgleason.org>.
- 104 <https://www.logopedie.nl/kennis/dysartrie/>. Creer S. et al. (2013). Building personalised synthetic voices for individuals with severe speech impairment. *Science Direct*. <https://www.sciencedirect.com/science/article/abs/pii/S0885230812000836>.
- 105 <https://www.filmacademie.ahk.nl/lichting/2020/projecten/deepfake-therapy/>. Zie ook: <https://www.beeldkompas.nl/kennisbank/deepfake-wat-is-het>.
- 106 Zhao, Y. (2018). Enabling people with visual impairments to navigate virtual reality with a haptic and auditory cane simulation. In Proceedings of the *2018 CHI conference on human factors in computing systems* (pp. 1-14).
- 107 Eng, K. et al. (2007). Cognitive virtual-reality based stroke rehabilitation. In *World Congress on Medical Physics and Biomedical Engineering 2006* (pp. 2839-2843). Springer, Berlin, Heidelberg.
- 108 Schwartz M. (2018). Who Killed the Kiev Protesters? *A 3-D Model Holds the Clues.* New York Times. <https://www.nytimes.com/2018/05/30/magazine/ukraine-protest-video.html>.
- 109 <https://www.terredeshommes.nl/nl/programmas/sweetie>.
- 110 Van der Hof, S., Georgieva, I., Schermer, B., & Koops, B. J. (Eds.). (2019). *Sweetie 2.0: Using artificial intelligence to fight webcam child sex tourism*. TMC Asser Press.
- 111 Baron K. (2019). Digital Doubles: The Deepfake Tech Nourishing New Wave Retail. *Forbes*. <https://www.forbes.com/sites/katiebaron/2019/07/29/digital-doubles-the-deepfake-tech-nourishing-new-wave-retail/?sh=c-4ce31f4cc7b>.
- 112 <https://www.charlietemple.com/nl-nl/virtuele-pas-kamer>.
- 113 Apparel Resources News-Desk (2020). Virtual fitting room market forecast to double by 2025: Report. *Apparel Resources*. <https://in.apparelresources.com/technology-news/retail-tech/virtual-fitting-room-market-forecast-double-2025-report/>.
- 114 <https://www.cereproc.com/en/jfkunsilenced>.
- 115 Chandler S. (2020). Why Deepfakes Ara A Net Positive For Humanity. *Forbes*. <https://www.forbes.com/sites/simonchandler/2020/03/09/why-deepfakes-are-a-net-positive-for-humanity/?sh=61c55efo2f84>.

- 116 <https://www.youtube.com/watch?v=8ooiOm-2sLw>.
- 117 <Deepempathy.mit.edu>.
- 118 Christopher N. (2020). We've Just Seen the First Use of Deepfakes in an Indian Election Campaign. Vice. <https://www.vice.com/en/article/jgedjb/the-first-use-of-deepfakes-in-indian-election-by-bjp>.
- 119 <https://www.reddithelp.com/hc/en-us/articles/360043075032>.
- 120 Fink D., Diamond S. (2020). Deepfakes:2020 and Beyond. *Law.com*. <https://www.law.com/therecorder/2020/09/03/deepfakes-2020-and-beyond/?slreturn=202100141010 12>.
- 121 Hern A. (2018). 'Deepfake' face-swap porn videos banned by Pornhub and Twitter. *The Guardian*. <https://www.theguardian.com/technology/2018/feb/07/twitter-pornhub-ban-deepfake-ai-face-swap-porn-videos-celebrities-gfycat-reddit>.
- 122 Bickert M. (2020). Enforcing Against Manipulated Media. *Facebook*. <https://about.fb.com/news/2020/01/enforcing-against-manipulated-media/>.
- 123 Sensity (2019). The State of Deepfakes: Landscape, Threats, and Impact. Medium. <https://medium.com/sensity/mapping-the-deepfake-landscape-27cb809e98bc>.
- 124 Melville K. (2019). The insidious rise of deepfake porn videos – and one woman who won't be silenced. ABC News. <https://www.abc.net.au/news/2019-08-30/deepfake-revenge-porn-noelle-martin-story-of-image-based-abuse/11437774>.
- 125 Zie echter ook. Thomas D. (2020). Deepfakes: a threat to democracy or just a bit of fun?. *BBC News*. <https://www.bbc.com/news/business-51204954>.
- 126 Schwartz O. (2018). You thought fake news was bad? Deep fakes are where thruth goes to die. *The Guardian*. <https://www.theguardian.com/technology/2018/nov/12/deep-fakes-fake-news-truth>.
- 127 Fagan K. (2018). A viral video that appeared to show Obama calling Trump a 'dips—' shows a disturbing new trend called 'deepfakes'. *Business Insider*. <https://www.businessinsider.nl/obama-deepfake-video-insulting-trump-2018-4?international=true&r=US>.
- 128 <https://www.youtube.com/watch?v=m7u-y90qUSw>.
- 129 <https://www.theverge.com/2020/4/28/21240488/jay-z-deepfakes-roc-nation-youtube-removed-ai-copyright-impersonation>.
- 130 Harbinja, E. (2017). Post-mortem privacy 2.0: theory, law, and technology. *International Review of Law, Computers & Technology, 31*(1), 26-42.
- 131 <https://slate.com/technology/2020/11/robert-kardashian-joaquin-oliver-deepfakes-death.html>.
- 132 Edwards J. (2019). A false rumor on WhatsApp started a run on a London bank. *Business Insider*. <https://www.businessinsider.nl/whatsapp-rumour-started-run-on-metro-bank-2019-5?international=true&r=US>.
- 133 <https://www.europol.europa.eu/sites/default/files/documents/malicious_uses_and_abuses_of_artificial_intelligence_europol.pdf>.
- 134 Stupp C. (2019). Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case. *The Wall Street Journal*. <https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402>.
- 135 The Observers (2019). Deepfake video of former Italian PM Matteo Renzi sparks debate in Italy. *The Observers*. <https://observers.france24.com/en/20191008-deepfake-video-former-italian-pm-matteo-renzi-sparks-debate-italy>.
- 136 <https://observers.france24.com/en/20191008-deepfake-video-former-italian-pm-matteo-renzi-sparks-debate-italy>.
- 137 Dobber, T. et al. (2020). Do (microtargeted) deepfakes have real effects on political attitudes?. *The International Journal of Press/Politics*, 1940161220944364.
- 138 Zie ook: <https://www.rathenau.nl/nl/digitale-samenleving/digitale-dreigingen-voor-de-democratie>.
- 139 Tweede Kamer, Brief heimelijke beïnvloeding van de publieke opinie door statelijke actoren, 8e1652c7-or1-3.0.
- 140 <https://www.motherjones.com/politics/2019/03/deepfake-gabon-ali-bongo/>.
- 141 BBC (2020). Australia demands China apologise for posting 'repugnant' fake image. *BBC News*. <https://www.bbc.com/news/world-australia-55126569>. In hoeverre het hier ging om een echte Deepfake, of een nepfoto gegener-

eerd door een minder geavanceerde techniek is niet met zekerheid vast komen te staan.

- 142 <https://www.europol.europa.eu/sites/default/files/documents/malicious_uses_and_abuses_of_artificial_intelligence_europol.pdf>.
- 143 Vaccari, C., & Chadwick, A. (2020). Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media+ Society, 6*(1), 2056305120903408.
- 144 Swerling G. (2020). Doctored aufio evidence used to damn father in custody battle. *The Telegraph*. <https://www.telegraph.co.uk/news/2020/01/31/deepfake-audio-used-custody-battle-lawyer-reveals-doctored-evidence/>.
- 145 <https://www.groene.nl/artikel/misogynie-als-politiek-wapen>.
- 146 <https://www.youtube.com/watch?v=B5eMz4JpYu0>. <https://www.youtube.com/watch?v=4b79yBzyRHs>.
- 147 McIntyre, Lee. Post-truth. MIt Press, 2018.
- 148 Higgins, K. (2016). Post-truth: a guide for the perplexed. *Nature News*, 540(7631), 9.
- 149 Suiter, J. (2016). Post-truth politics. *Political Insight, 7*(3), 25-27.
- 150 Pariser, E. (2011). *The filter bubble: How the new personalized web is changing what we read and how we think.* Penguin.
- 151 <https://edition.cnn.com/2021/02/04/politics/2020-election-donald-trump-voter-fraud/index.html>.
- 152 Radio uitzending NPO radio 1.
- 153 Zie verder: Huijstee, M. et al. (July 2021). *Tackling deepfakes in European policy.* Brussels: European Parliament Research Service, 116 pp.

Origineel

Filter

Gecorrigeerd

Aangepast

Geanceneerd

Deepfake

Niet elk synthetisch beeld is een deepfake. Er bestaan velerlei gradaties van beeld- of audiobewerking. Welke vormen van manipulatie zijn toegestaan binnen welke context? En hoe moeten we dit reguleren?

3. Materieelrecht

Dit hoofdstuk bespreekt het huidige juridische kader ten aanzien van deepfakes en brengt in kaart waar mogelijke lacunes zijn. Daarbij worden vier rechtsgebieden besproken: het strafrecht (paragraaf 3.1), het gegevensbeschermingsrecht (paragraaf 3.2), de vrijheid van meningsuiting en het recht op eer en goede naam (paragraaf 3.3), het portretrecht (paragraaf 3.4) en zal kort gewag worden gemaakt van de regels die in de nu door Europese Commissie voorgestelde AI Regulation staan omtrent deepfakes (paragraaf 3.5). Vervolgens zal worden besproken het onrechtmatige-daadsrecht (paragraaf 3.6). Ook zal een korte conclusie worden geboden (paragraaf 3.7).

3.1 Strafrecht

3.1.1 Inleiding

In deze paragraaf worden deepfakes geanalyseerd in relatie tot de huidige strafbepalingen in het Wetboek van Strafrecht. In de bespreking worden drie clusters onderscheiden van bepalingen die verschillende belangen beogen te beschermen: 1) financieel-economische belangen, 2) privacy- en reputatiebelangen en 3) gegevensgerelateerde belangen. De eerste twee clusters komen grotendeels overeen met het in paragraaf 2.5 gemaakte onderscheid tussen deepfakes als middel tot iets anders (vorm) en deepfakes als doel in zichzelf (inhoud). Bij deepfakes als vehikel tot een (ander) strafbaar feit spelen vaak financieel-economische motieven een rol; bij deepfakes als doel in zichzelf gaat het vaak om inbreuken op de privacy of reputatie van anderen. De indeling middel/doel is echter minder goed toepasbaar op strafbepalingen die ook relevant zijn om strafwaardige deepfakes te bestrijden, met name de computercriminaliteitsbepalingen betreffende gegevensaantasting. Daarom wordt een hoofdindeling gekozen die de nadruk legt op de verschillende te beschermen belangen.

Bij de analyse hebben is gekozen voor een combinatie van breedte en diepte. Aangezien het om een inventarisatie gaat van mogelijke lacunes in de wet, is het belangrijk om alle mogelijk relevante bepalingen in ogenschouw te nemen; daarom wordt een relatief grote hoeveelheid bepalingen genoemd. Veel daarvan zijn echter alleen op heel specifieke situaties van toepassing (bijv. koersmanipulatie) of zijn evident strafbaar (kinderpornografie); die bepalingen zullen daarom alleen kort worden genoemd. Andere strafbepalingen zijn echter relevanter, hetzij omdat zij van toepassing zouden kunnen zijn op een breed scala aan deepfakes (bijv. gegevensaantasting) of omdat zij belangrijke interpretatievragen oproepen (bijv. identiteitsfraude of het wederrechtelijk maken van seksuele afbeeldingen). Die bepalingen worden dan ook uitgebreider geanalyseerd.

Bij de onderstaande analyse moet steeds een algemeen aspect voor ogen worden gehouden. Dit betreft de vraag of een bepaalde handeling met deepfakes als **onrechtmatig** kan worden gekwalificeerd. Onrechtmatigheid is immers vaak (maar niet altijd) in de vorm van het bestanddeel ‘wederrechtelijk’ een onderdeel van de delictsomschrijving, dat dus moet worden bewezen. Bij de kwalificatie van een deepfake als al dan niet onrechtmatig speelt het spanningsveld met de vrijheid van meningsuiting een belangrijke rol. De afweging tussen de verschillende belangen hangt van allerlei contextfactoren af en kan daarom alleen casuïstisch – voor elk geval afzonderlijk – worden gemaakt. Vooral bij uitingsdelicten als

laster, belediging en discriminatie is de grens tussen een strafbare uiting en een toelaatbare uitoefening van de vrije meningsuiting dun en moeilijk bepaalbaar. Dit is evenwel niet specifiek voor deepfakes, maar geldt voor alle manieren waarop uitingsdelicten kunnen worden gepleegd. Daarom zal niet nader worden ingegaan op dit aspect; de rechtspraktijk zal de interpretatie van deepfakes als al dan niet wederrechtelijk kunnen inpassen in de jurisprudentieontwikkeling, zoals dat steeds is gebeurd bij nieuwe contexten of nieuwe middelen voor het plegen van de desbetreffende delicten.

3.1.2 Strafbepalingen ter bescherming van financieel-economische belangen

Het eerste cluster betreft strafbepalingen die primair ***financiële en economische belangen*** beogen te beschermen, zoals eigendomsrechten en vertrouwen in het (economische) rechtsverkeer. Bij deze delicten is het opzet van deepfakemakers vooral gericht op financieel gewin; deepfakes zijn hier dus een nieuw middel om reeds lang bestaande financieel-economische delicten te plegen.

3.1.2.1 Oplichting (art. 326 Sr)

Deepfakes zijn bijzonder geschikt om mensen op te lichten, doordat zij immers de indruk wekken een boodschap over te brengen van een bepaalde persoon, terwijl die persoon nooit een dergelijke communicatie heeft gedaan. De laatste jaren neemt oplichting in de vorm van WhatsAppfraude (ook wel vriend-in-noodfraude genoemd)[154] sterk toe. Hierbij doet iemand zich, via een gehackt of vals account, voor als een familielid of vriend die in nood zit en dringend geld nodig heeft. Deepfakes kunnen het bedrog aanzienlijk overtuigender maken, door het levensecht nabootsen van iemands stem of uiterlijk (al is het real-time genereren van zulke deepfakes technisch nog niet enorm ver gevorderd). Op dezelfde manier versterken deepfakes ook de mogelijkheid van CEO-fraude, een andere vorm van oplichting die een hoge vlucht heeft genomen. Hierbij doet iemand zich voor tegenover een financieel-administratief medewerker als de baas (CEO of CFO) die vraagt om met spoed een groot bedrag over te maken.[155] Hoewel deze vorm van fraude vooral slaagt bij medewerkers die de baas niet persoonlijk kennen, kan een deepfake – zeker bij publieke bekende CEO's – wel helpen om het verzoek overtuigender te maken. Naast vriend-in-noodfraude en CEO-fraude zijn er natuurlijk vele andere vormen om mensen op te lichten, waarbij deepfakes kunnen helpen om mensen te misleiden tot het afgeven van hun geld of goed.

Deepfakes zijn dus een belangrijk nieuw hulpmiddel voor het plegen van oplichting (zoals de juridische benaming voor fraude luidt). Artikel 326 lid 1 Sr stelt oplichting als volgt strafbaar:

> *'Hij die, met het oogmerk om zich of een ander wederrechtelijk te bevoordelen, hetzij door het aannemen van een valse naam of van een valse hoedanigheid, hetzij door listige kunstgrepen, hetzij door een samenweefsel van verdichtsels, iemand beweegt tot de afgifte van enig goed, tot het verlenen van een dienst, tot het ter beschikking stellen van gegevens, tot het aangaan van een schuld of tot het teniet doen van een inschuld, wordt, als schuldig aan oplichting, gestraft met gevangenisstraf van ten hoogste vier jaren of geldboete van de vijfde categorie.'*

Relevant voor dit rapport is het bestanddeel van de modus operandi: het bedrog moet plaatsvinden door een valse naam of

hoedanigheid, listige kunstgrepen of een samenweefsel van verdichtsels. Deepfakes zullen bij vriend-in-nood- en CEO-fraude vallen onder het aannemen van een valse naam of hoedanigheid, bijvoorbeeld de zoon-in-nood of de baas die dringend geld overgemaakt willen hebben. In andere situaties, bijvoorbeeld als de deepfake niet een bekende voorstelt, of zelfs een niet-bestaande persoon betreft, zal de deepfake eenvoudig kunnen worden gekwalificeerd als een listige kunstgreep (geen gekke omschrijving van het verschijnsel deepfakes in het algemeen!) of als een onderdeel van een samenweefsel van verdichtsels waarmee iemand wordt misleid om een goed af te staan. Deze strafbepaling is dus goed toegesneden op deepfakes en kent geen lacunes of onduidelijkheden.

Hetzelfde geldt voor specifieke vormen van oplichting, die ook technologie-neutraal zijn geformuleerd en daarom ook op deepfake-bedrog van toepassing kunnen zijn. Denk bijvoorbeeld aan deepfakes die worden gebruikt voor acquisitiefraude (art. 326d Sr) of verzekeringsfraude (art. 327 Sr).

3.1.2.2 Afpersing en afdreiging (art. 317-318 Sr)

Afpersing wordt gepleegd wanneer iemand door geweld of dreiging van geweld een ander dwingt tot afgifte van een goed of gegevens (art. 317 Sr). Een deepfakefilmpje of afbeelding is zelf geen geweld, maar kan natuurlijk wel dreiging van geweld bevatten. Het geweld hoeft niet tegen de afgeperste zelf gericht te zijn;[156] denkbaar is bijvoorbeeld dat iemand een bericht krijgt met het 'verzoek' om €10.000 aan bitcoins te overhandigen, met bijvoeging van een filmpje van een onthoofding waarbij het slachtoffer een deepfake is van de zoon van de geadresseerde. Deepfaketechnologie maakt zulke dreiging directer, maar verschilt niet wezenlijk van gewone of gefotoshopte afbeeldingen waar dreiging van uitgaat. Een onthoofdingsdeepfake is evenzeer dreigen met geweld als het versturen van een brief met een kogel en een foto van de zoon. Deepfakes zullen in dit opzicht geen interpretatieproblemen of lacunes opleveren voor afpersing.

Het strafrecht kent voorts het aanpalende delict van afdreiging. Artikel 318 lid 1 Sr luidt:

> *'Hij die, met het oogmerk om zich of een ander wederrechtelijk te bevoordelen, door bedreiging met smaad, smaadschrift of openbaring van een geheim iemand dwingt hetzij tot de afgifte van enig goed dat geheel of ten dele aan deze of aan een derde toebehoort, hetzij tot het aangaan van een schuld of het teniet doen van een inschuld, hetzij tot het ter beschikking stellen van gegevens, wordt als schuldig aan afdreiging, gestraft met gevangenisstraf van ten hoogste vier jaren of geldboete van de vijfde categorie.'*

Afdreiging behelst hetzelfde als afpersing, behalve dat de dreiging nu niet uitgaat van geweld maar van – vooral – reputatieschade door bepaalde feiten bekend te maken of te insinueren. Voor dreiging met openbaring van een geheim geldt hetzelfde als voor geweld: dat kan door een deepfake – bijvoorbeeld een filmpje van twee zoenende jongens waarin een nog in de kast zittende gereformeerde jongen is gedeepfaket – maar dat verschilt niet wezenlijk van een e-mail waarin gedreigd wordt de vader te vertellen dat de zoon in een homokroeg is gesignaleerd. Beide vormen zijn eenvoudig te kwalificeren als het bestanddeel 'bedreiging met openbaring van een geheim'.

Voor bedreiging met smaad of smaadschrift (aanranden van iemands goede naam door beschuldiging van een bepaald feit, zie par. 3.1.3) geldt hetzelfde. Smaad(schrift) is weliswaar iets lastiger te herkennen dan geweld of openbaring van een geheim; de grens tussen een strafbare smadelijke uiting en een toelaatbare bekendmaking van een bepaald feit is dun en vergt specialistische kennis van de jurisprudentie. Maar ook dat is niet specifiek voor deepfakes: de dreiging van ruchtbaarheid aan een bepaald feit – bijvoorbeeld dat de arts die voor de Hartstichting een gezonde levensstijl aanprijst, feitelijk een kettingroker is – kan evenzeer worden gepleegd middels een tekst of een gemanipuleerde foto als via een deepfake.

Evenals bij oplichting blijkt deepfake dus simpelweg een nieuwe techniek of methode om afpersing of afdreiging te plegen, die niet tot interpretatieproblemen of lacunes leidt vanwege de techniek-neutrale formulering van de modus operandi.

Vergelijkbare overwegingen gelden overigens voor het aan afpersing en afdreiging verwante delict dwang (art. 284 Sr), waar een bedreiging met geweld of met smaad(schrift) plaatsvindt niet om geld te verkrijgen maar om een andere handeling (of het nalaten van een handeling) af te dwingen. Dit valt niet onder de financieel-economische delicten maar onder inbreuken op de autonomie, maar dat doet niet af aan de toepasbaarheid op deepfake-bedreigingen.

3.1.2.3 Identiteitsfraude (art. 231a-231b Sr)

Een specifiek strafbaar gestelde vorm van oplichting is identiteitsfraude, dat wil zeggen situaties waarin iemands identiteit wordt misbruikt om fraude te plegen. De strafbaarstelling van identiteitsmisbruik is echter breder dan oplichting: het gaat om alle situaties waarin iemand biometrische gegevens misbruikt voor identificatiedoeleinden (art. 231a Sr) of andere identificerende gegevens misbruikt waardoor nadeel kan ontstaan (art. 231b Sr).[157] Deepfakes bevatten veelal identificerende gegevens van iemand, zoals stem en/of gelaat, zodat deze strafbepalingen een belangrijk middel kunnen zijn om strafwaardige deepfakes te bestrijden.

De relevante delen van artikel 231a Sr – biometrische identiteitsfraude – luiden als volgt:

> *'1. Hij die biometrische kenmerken of biometrische persoonsgegevens valselijk opmaakt of vervalst met het oogmerk om deze als echt en onvervalst te gebruiken of te doen gebruiken in gevallen waarin die kenmerken of persoonsgegevens worden gebruikt voor het vaststellen van iemands identiteit, teneinde zijn identiteit te verhelen of de identiteit van een ander te verhelen of misbruiken, wordt gestraft met gevangenisstraf van ten hoogste zes jaren of geldboete van de vijfde categorie.*
>
> *2. Met dezelfde straf wordt gestraft hij die in gevallen waarin biometrische kenmerken of biometrische persoonsgegevens worden gebruikt voor het vaststellen van iemands identiteit, opzettelijk gebruik maakt van valse of vervalste biometrische kenmerken of biometrische persoonsgegevens als waren deze echt en onvervalst met het oogmerk om zijn identiteit te verhelen of de identiteit van een ander te misbruiken of opzettelijk gebruik maakt van biometrische kenmerken of biometrische persoonsgegevens van een ander met het oogmerk om de verdenking van een strafbaar feit op de ander of niet op hem te doen ontstaan (...).'*

Lid 1 stelt aldus de vervalsing van biometrische gegevens strafbaar, en lid 2 het gebruik van zulke vervalste biometrie, als dit gebeurt met het oogmerk een biometrische identiteitscontrole te misleiden. De meest voorkomende vorm hiervan is financiële identiteitsfraude, waarbij iemands identificerende gegevens worden misbruikt voor oplichting. Het kan echter ook de vorm aannemen van strafrechtelijke identiteitsfraude, waarop het laatste zinsdeel van artikel 231a lid 2 Sr ziet: het misbruiken van biometrie om aan verdenking van een strafbaar feit te ontkomen. Dit kan bijvoorbeeld door de vingerafdruk van iemand anders te gebruiken om een beveiligd gebouw binnen te komen om een moord te plegen, maar ook door de haren van iemand anders rond te strooien op een plaats delict; in zo'n situatie wordt immers DNA-materiaal gebruikt voor het vaststellen van de identiteit van mogelijke daders.[158] Deepfaketechnologie zou gebruikt kunnen worden om de schuld aan een strafbaar feit te verhullen. Denk bijvoorbeeld een video waarin gesuggereerd wordt dat iemand anders doodsbedreigingen uit aan het adres van een beoogd moordslachtoffer, waardoor de verdenking op die persoon wordt gericht. Ook zou iemand een video kunnen maken waarin een beoogd moordslachtoffer zelfmoordgedachten lijkt te uiten, zodat de moord onterecht als zelfmoord wordt beschouwd.

Overige vormen van identiteitsfraude vallen onder artikel 231b Sr:

> *'Hij die opzettelijk en wederrechtelijk identificerende persoonsgegevens, niet zijnde biometrische persoonsgegevens, van een ander gebruikt met het oogmerk om zijn identiteit te verhelen of de identiteit van de ander te verhelen of misbruiken, waardoor uit dat gebruik enig nadeel kan ontstaan, wordt gestraft met een gevangenisstraf van ten hoogste vijf jaren of geldboete van de vijfde categorie.'*

Op het oog bieden deze bepalingen, en met name de laatste, een ruime mogelijkheid om kwalijke deepfakes aan te pakken. Bij deepfakes, althans deepfakes van bestaande personen, worden immers iemands persoonsgegevens gebuikt op een manier die vaak kan worden gekwalificeerd als misbruik van diens identiteit, terwijl ook snel 'enig nadeel' kan ontstaan. Dat nadeel kan volgens de wetsgeschiedenis vele vormen aannemen, zoals 'direct financieel nadeel, reputatieschade of schade door het vervuilen van (overheids)databases'.[159]

Bij nadere bestudering lijken deepfakes echter te vallen in een gat tussen beide bepalingen. Artikel 231a Sr is ingevoerd om met name fraude met biometrie in identiteitsbewijzen tegen te gaan.[160] De bepaling is dan ook toegesneden op situaties waarin biometrie wordt gebruikt voor de vaststelling van iemands identiteit. Dat hoeft overigens niet per se een paspoort of rijbewijs te betreffen – het kan ook gaan om bijvoorbeeld een toegangscontrole tot een gebouw middels gezichtsherkenning, of een toegangscontrole tot een computer via een vingerafdruksensor. In de parlementaire behandeling is later, ter aanvulling van artikel 231 Sr (fraude met identiteitsbewijzen) en artikel 231a Sr, artikel 231b Sr toegevoegd, omdat 'het aantal slachtoffers van fraude met identificerende persoonsgegevens de afgelopen jaren explosief toeneemt.'[161]

Bij de redactie van artikel 231b Sr is echter gekozen – vermoedelijk om uitdrukkelijk overlap met artikel 231a Sr te voorkomen – voor misbruik van 'identificerende persoonsgegevens, *niet*

zijnde biometrische persoonsgegevens' (cursivering toegevoegd). Met deze formulering wordt gedoeld op 'alle gegevens waarmee een persoon kan worden geïdentificeerd, zoals (combinaties van) naam, adres, telefoonnummer, accounts, handles, nicknames etc. etc.'[162] Hoewel de opsomming in de bijzin niet uitputtend is, is duidelijk gedacht aan beschrijvende persoonsgegevens, zoals namen en klantnummers. Hoewel de hoofdzin in de toelichting spreekt van 'alle gegevens waarmee een persoon kan worden geïdentificeerd', waaronder dus ook biometrische gegevens vallen, sluit de wettekst zelf biometrische gegevens expliciet uit. Daarom zal de grammaticale interpretatie moeten prevaleren boven een teleologische interpretatie: artikel 231b Sr is niet van toepassing op 'biometrische persoonsgegevens'. Deze term wordt in de Memorie van Toelichting als volgt omschreven:

> *'Onder biometrische kenmerken vallen de lichamelijke kenmerken en gedragskenmerken van een persoon die gebruikt kunnen worden voor het vaststellen of controleren van iemands identiteit en die voor ieder persoon binnen bepaalde grenzen uniek zijn. Bij lichaamskenmerken gaat het om het gezicht, de ogen, waaronder de iris, de vingers, de handpalmen, de aders en het DNA-materiaal en bij gedragskenmerken om de manier van schrijven, lopen en spreken. De gegevens die naar aanleiding van de biometrische kenmerken worden verkregen, worden in artikel 231a Sr aangeduid als biometrische persoonsgegevens.'*[163]

Als voorbeelden worden genoemd 'iemands vingerafdrukken, gelaatsfoto, stemopname, handtekening of DNA-profiel.'[164] Nu spelen bij deepfakes juist het gelaat en de stem een belangrijke rol, meer dan beschrijvende persoonsgegevens als namen of klantnummers. Dit betekent dat deepfakevideo's of -audiobestanden alleen onder artikel 231b Sr zullen vallen indien de bestanden naast iemands uiterlijk of stem ook diens (bij)naam of een ander beschrijvend identiteitskenmerk vermelden. De strafbaarheid zit hem dan in het misbruik van de aanduiding van de persoon, niet in het misbruik van het uiterlijk of de stem. De kern van de meeste deepfakes is echter juist dit laatste: het gebruiken van iemands herkenbare uiterlijk, stem of gedrag, om de indruk te wekken dat het om die persoon gaat.

Voor de strafbaarheid van deepfakes als zodanig betekent dit dat identiteitsfraude eigenlijk vooral toepasbaar is wanneer deepfakes worden gebruikt om de schuld aan een misdrijf te verhullen of om biometrische identiteitscontroles te omzeilen; dan is artikel 231a Sr toepasbaar. Het omzeilen van identiteitscontroles door middel van deepfakes zal zelden voorkomen; er zijn in deze studie weinig voorbeelden naar voren gekomen van deepfakes om onrechtmatig een gebouw of evenement binnen te komen of anderszins identiteitscontroleurs te foppen. Artikel 231b Sr zal niet kunnen worden ingeroepen om deepfakes als zodanig te bestrijden waarbij zonder iemands toestemming diens beeld of stem wordt gebruikt en 'enig nadeel' kan ontstaan, behalve in gevallen waarin ook expliciet de (bij)naam of andere beschrijvende aanduiding van de persoon wordt vermeld. Of dat een ernstige lacune is, hangt af van de vraag of andere strafbaarstellingen voldoende dergelijke situaties afdekken. De conclusie van deze paragraaf (3.1.5) komt hierop terug.

3.1.2.4 Overige bepalingen

In bepaalde gevallen kunnen deepfakes ook worden gebruikt als middel om andere strafbare

feiten te plegen. Theoretisch zou iemand bijvoorbeeld een bedrijfsgeheim van Facebook openbaar kunnen maken (zonder dat dit nodig is voor het publiek belang) door Mark Zuckerberg dit in een deepfakefilmpje te laten uitspreken; het maakt voor de strafbaarheid onder artikel 273 Sr (schending van bedrijfsgeheimen) niet uit of dit gebeurt door een deepfakefilmpje of een andere publicatievorm. Hetzelfde geldt voor bepalingen die meer algemene financieel-economische belangen beschermen, zoals oneerlijke mededinging (art. 328bis Sr) en koersmanipulatie (art. 334 Sr); Deepfakes zouden een aantrekkelijk middel kunnen zijn voor dergelijke doeleinden (bijvoorbeeld een deepfake waarin Zuckerberg aankondigt te stoppen, om de aandelenkoers van Facebook te manipuleren). Aangezien die delicten allemaal technologie-neutraal zijn geformuleerd, zijn geen belemmeringen bij de toepasbaarheid op deepfakes te voorzien.

Verder kunnen soms andere strafbepalingen worden gebruikt om bepaalde toepassingen van deepfakes te bestrijden. Een voorbeeld is valsheid in geschrifte, artikel 225 Sr. Dit betreft het valselijk opmaken van een geschrift met het oogmerk dit als echt te (laten) gebruiken. In de jurisprudentie is reeds lang aangenomen dat 'geschrift' niet alleen slaat op papieren documenten, maar ook op elektronische bestanden;[165] ook filmpjes en geluidsopnames kunnen dus als geschrift in de zin van artikel 225 Sr worden gekwalificeerd. Voorwaarde voor valsheid in geschrifte is wel dat het geschrift een bewijsbestemming heeft in het rechtsverkeer. Dat zal voor de overgrote meerderheid van deepfakes niet het geval zijn. Uitzonderingen zijn echter denkbaar, bijvoorbeeld wanneer iemand in een vechtscheidingsprocedure een gedeepfaket filmpje inbrengt waarin de echtgenote de dochter een pak slaag geeft. Ook zouden soms deepfakes die niet direct gemaakt zijn om een zaak juridisch te flessen, later wel gebruikt kunnen worden als beweerd bewijs van een bepaald feit, bijvoorbeeld een voor de grap gemaakte porno-deepfake die in de vechtscheiding wordt ingebracht om de partner in diskrediet te brengen. De toepasbaarheid van valsheid in geschrifte hangt in zulke gevallen vooral af van de interpretatie van de bewijsbestemming; daarin verschillen deepfakes niet van bijvoorbeeld vervalste e-mails of Word-documenten. Wanneer valsheid in geschrifte niet kan worden bewezen, zijn in bijzondere gevallen ook andere valsheidsdelicten denkbaar, zoals het verstrekken van valse gegevens in het kader van subsidieverlening (art. 227a Sr, of als dit zonder opzet gebeurt, de overtreding van art. 447c Sr).

3.1.3 Strafbepalingen ter bescherming van privacy en reputatie

Het tweede cluster betreft strafbepalingen die primair **privacybelangen** beogen te beschermen. Onder privacybelangen kunnen ook, gezien de jurisprudentie over artikel 8 EVRM, **reputatie** en bescherming van de identiteitsconstructie van individuen worden begrepen.[166] Bij deze delicten is het opzet van deepfakemakers veelal gericht op het beschadigen van iemands reputatie, uit discriminatoire, pest- of wraakmotieven. In die zin zijn zulke deepfakes op zich ook een nieuw middel om reeds lang bestaande strafbare feiten te plegen, en geen doel op zich. Maar dergelijke deepfakes kunnen ook als een doel op zichzelf fungeren, wanneer het feit wordt gepleegd niet zozeer omdat iemand toch al de bedoeling had om te discrimineren of beledigen en daarvoor deepfakes kiest als een toevallig voorhanden hulpmiddel, maar waar juist de functionaliteit van deepfakes iemand

op het idee brengt om iemands reputatie aan te tasten. De grens tussen middel en doel is niet scherp te trekken hier. Wat daar ook van zij, de analyse van privacygerelateerde bepalingen wordt begonnen met enkele overkoepelende opmerkingen, alvorens de uiteenlopende strafbaarstellingen langs te lopen.

3.1.3.1 Algemene opmerkingen

Misdrijven vereisen over het algemeen opzet; bij de nodige misdrijven is opzet een expliciet bestanddeel, bij andere alleen impliciet.[167] De ondergrens van strafrechtelijk opzet ligt tamelijk laag: ook het bewust de kans aanvaarden dat een bepaald gevolg optreedt, valt eronder, ook als de dader niet per se dat gevolg beoogd heeft (zogeheten voorwaardelijk opzet). Deepfakes zullen altijd (min of meer) bewust worden gemaakt, maar niet per se met een bepaalde misdadige of schadelijke bedoeling. Deepfakes kunnen ook 'voor de grap' of uit balorigheid worden gemaakt, zonder opzet op reputatieschade, maar wel met reputatieschade of psychisch leed als gevolg. Ook kunnen sommige satirisch bedoelde deepfakes op een slachtoffer discriminerend of smadelijk overkomen, zonder dat ze zodanig bedoeld zijn. De interpretatie van opzet bij uitingsdelicten kan dan ook ingewikkeld zijn: het vergt een sterk contextueel afhankelijke analyse of iemand had moeten begrijpen (en dus bewust de kans aanvaardde) dat een bepaalde uiting discriminerend, smadelijk of beledigend is. Dat is altijd zo bij uitingsdelicten; mogelijk maken deepfakes de interpretatie echter nog wat lastiger, omdat er nog relatief weinig ervaring mee is in het maatschappelijk verkeer, zodat zowel deepfakemakers en verspreiders als slachtoffers nog niet goed kunnen inschatten hoe bepaalde deepfakes bedoeld zijn of hoe zij overkomen op anderen. Dat betreft geen lacune of tekortkoming in de wet, maar een uitdaging voor Openbaar Ministerie en rechterlijke macht om via jurisprudentieontwikkeling hierin piketpalen te slaan.

Een ander algemeen aspect van veel uitingsdelicten zijn de begrippen 'openbaar' in de discriminatiebepalingen en 'openlijk' in de strafbaarstelling van smaad en laster. In de context van sociale media en online platforms is het lang niet altijd duidelijk onder wie een bepaalde uiting (potentieel) wordt verspreid. Van algemeen toegankelijke Twitter- en YouTube-accounts is duidelijk dat ze openbaar zijn, maar bij een besloten Facebook- of WhatsApp-groep is het minder duidelijk. Aangezien de vraag waar de grens ligt tussen openbaar en niet-openbaar niet specifiek is voor deepfakes, maar voor alle typen uitingen in een online context, wordt hier niet nader op ingegaan.

3.1.3.2 Discriminatie (art. 137c e.v. Sr)

Belediging van bevolkingsgroepen en discriminatie zijn strafbaar gesteld in de artikelen 137c tot en met 137g Sr. Artikel 90quater Sr definieert discriminatie als

> *'elke vorm van onderscheid, elke uitsluiting, beperking of voorkeur, die ten doel heeft of ten gevolge kan hebben dat de erkenning, het genot of de uitoefening op voet van gelijkheid van de rechten van de mens en de fundamentele vrijheden op politiek, economisch, sociaal of cultureel terrein of op andere terreinen van het maatschappelijk leven, wordt teniet gedaan of aangetast.'*

Strafbaar zijn gesteld het in het openbaar beledigen van een groep mensen wegens onder andere ras, godsdienst of seksuele gerichtheid (art. 137c Sr), het in het openbaar aanzetten

tot haat, discriminatie of geweld tegen mensen wegens dergelijke karakteristieken (art. 137d Sr), het openbaar maken van beledigende of wegens dergelijke karakteristieken tot haat of discriminatie aanzettende uitingen (art. 137e Sr), het faciliteren van discriminatie (art. 137f Sr) en beroeps- of bedrijfsmatige rassendiscriminatie (art. 137g Sr).

De uitingsvorm van deze delicten is volgens de delictsomschrijvingen mondeling of bij geschrift of afbeelding (art. 137c/d Sr), dan wel een uitlating of een voorwerp dat een uitlating bevat (art. 137e). Deepfakes kunnen eenvoudig als zodanig worden gekwalificeerd, namelijk als mondeling (bij deepfake-geluidsopnames, voor zover deze worden afgespeeld), als afbeelding (bij deepfake-filmpjes), als uitlating of voorwerp bevattende een uitlating (waaronder ook geluids- en beeldopnames vallen).[168] Ook hier zijn de delictsomschrijvingen dus voldoende techniek-neutraal geformuleerd.

Of een deepfake-uiting daadwerkelijk discriminatie oplevert, zal vooral afhangen van de vraag of de uiting in de desbetreffende context beledigend is of tot haat of discriminatie aanzet. Dat is bij deze delicten altijd een complexe vraag, die op basis van in de jurisprudentie ontwikkelde factoren moet worden beantwoord. Het gegeven dat een filmpje of geluidsopname een deepfake is, maakt die complexe vraag niet anders. Wel kan dit gegeven een rol spelen in de beoordeling, bijvoorbeeld bij een filmpje van een cabaretier waarin uitlatingen over transgenders op het randje van de toelaatbaarheid zijn; vanwege de artistieke of 'humoristische exceptie' zou de cabaretier het voordeel van de twijfel moeten worden gegund. Als het filmpje echter een deepfake blijkt te zijn, zou de exceptie-grondslag vervallen en het juist wel als strafbaar kunnen worden gekwalificeerd, ook omdat de uiting in een misleidende vorm is gegoten waarbij de functie van de cabaretier is misbruikt. Daarentegen zou, afhankelijk van de omstandigheden, de maker van zo'n filmpje zich ook kunnen beroepen op de artistieke of humoristische exceptie,[169] als het filmpje bijvoorbeeld vrij duidelijk een deepfake is en volgens de maker of openbaarmaker juist grappig bedoeld is. Of wellicht dient zo'n filmpje het publiek belang, als het bedoeld is om een debat aan te jagen over een cabaretier die volgens de maker of openbaarmaker te vaak over de schreef gaat met discriminerende grappen. De kwalificatie van ogenschijnlijk discriminerende deepfake-uitingen zal aldus vaak een complexe afweging opleveren, maar daarin verschilt deepfake niet van andere vormen waarin discriminerende uitingen worden gegoten.

3.1.3.3 Smaad, smaadschrift, laster en belediging (art. 261, 262 en 266 Sr)

Een cluster delicten onder de noemer 'Belediging' (titel XVI van Boek 2) stelt diverse vormen van aantasting van iemands eer of reputatie strafbaar. Smaad betreft het opzettelijk aanranden van iemands eer of goede naam, door telastlegging van een bepaald feit, met het kennelijke doel daaraan ruchtbaarheid te geven (art. 261 Sr). Wanneer dit gebeurt door een geschrift of afbeelding te verspreiden of tentoon te stellen, is er sprake van smaadschrift (art. 261 lid 2 Sr). Laster is een bepaalde vorm van smaad(schrift), namelijk wanneer iemand weet dat het feit waarvan hij of zij iemand beschuldigt, onwaar is (art. 262 Sr). Smaad(schrift) (en dus ook laster)[170] kunnen ook ten aanzien van overledenen worden gepleegd (art. 270 Sr).

Deepfakefilmpjes kunnen als afbeelding onder smaadschrift (en dus ook laster) vallen.[171] Ook het afspelen van deepfake-geluidopnames kan eronder vallen, aangezien smaadschrift ook omvat 'geschriften waarvan de inhoud openlijk ten gehore wordt gebracht' (art. 261 lid 2 Sr); de wetgever doelde daarmee op grammofoonplaten die openlijk werden afgespeeld, maar ook gegevensdragers die via radio, televisie of internet worden afgespeeld vallen daaronder.[172] Zo kunnen bijvoorbeeld op internet geplaatste seksuele deepfakefilmpjes smaadschrift opleveren, aangezien ze de gedeepfakete persoon in verband brengen met seksuele handelingen en dit de eer of goede naam van die persoon kan aantasten.[173] Het (deep)linken naar zo'n filmpje kan daarbij het verspreiden van beledigende geschriften opleveren, mits de (deep)linker zich bewust is van de smadelijke inhoud (art. 271 Sr).[174] Een ander voorbeeld is het hierboven bij valsheid in geschrifte genoemde gedeepfakete filmpje waarin het lijkt alsof de echtgenote de dochter een pak slaag geeft; dat kan laster opleveren indien de partner dit filmpje in een Instagramgroep verspreidt met het kennelijke doel om ruchtbaarheid aan de vermeende kindermishandeling te geven, of lasterlijke aanklacht (art. 268 Sr) als de partner met dit filmpje aangifte doet van kindermishandeling.

Eenvoudige belediging betreft elke opzettelijke belediging die niet onder smaad(schrift) of laster valt (art. 266 Sr). Dit kan worden gepleegd via geschriften of afbeeldingen in het openbaar, maar ook door een geschrift of afbeelding (rechtstreeks of indirect) aan de beledigde persoon toe te zenden, bijvoorbeeld via email.[175] Voor belediging is niet per se het beschuldigen van een bepaald feit vereist; voldoende is dat een uiting iemands eer of goede naam aantast, wat het geval is 'indien de uiting de strekking heeft iemand in een ongunstig daglicht te plaatsen'.[176] Zo kan een via sociale media verspreid deepfakefilmpje waarin iemand de burgemeester uitscheldt voor klootzak, belediging opleveren (het deepfakekarakter doet daarbij weinig ter zake) – al hangt het af van de omstandigheden of het scheldwoord in die context onnodig grievend is. Een aan iemands benedenbuurvrouw verstuurd deepfakefilmpje waarin de bovenbuurman een enthousiast pleidooi lijkt te houden voor de platte-aarde-theorie, zou belediging kunnen opleveren, aangezien de gedeepfakete persoon hiermee vermoedelijk in een ongunstig daglicht wordt geplaatst; voorwaarde is wel dat men redelijkerwijs kon verwachten dat de benedenbuurvrouw het filmpje zou laten zien, of de inhoud doorvertellen, aan de buurman. Voor alle genoemde gevallen geldt overigens als voorwaarde dat de beledigde een klacht indient (art. 269 Sr).

Evenals bij discriminatie vergt de vraag of een bepaalde uiting smadelijk, lasterlijk of beledigend is een complexe afweging van allerlei contextuele factoren, zeker omdat de bepalingen expliciete uitzonderingsgronden formuleren voor uitingen die in het publiek belang worden gedaan (art. 261 lid 3 Sr, art. 266 lid 2 Sr). Dat is bij deepfake-uitingen niet anders dan bij 'echte' uitingen.

3.1.3.4 Wraakporno en het wederrechtelijk maken van seksuele afbeeldingen (art. 139h Sr)

Sinds 2020 is wraakporno – het verspreiden van seksuele afbeeldingen van iemand zonder diens toestemming – expliciet strafbaar gesteld. Wraakporno kan ook onder andere strafbaarstellingen vallen, zoals smaad en belediging, maar om eventuele lacunes in de wet te dichten heeft de wetgever in artikel 139h

lid 2 onder b Sr een expliciete strafbaarstelling ingevoerd voor degene 'die van een persoon een afbeelding van seksuele aard openbaar maakt, terwijl hij weet dat die openbaarmaking nadelig voor die persoon kan zijn'. Een afbeelding is een foto, video of live streamingbeelden; deze zijn van seksuele aard als de afbeelding 'een zodanig intiem seksueel karakter heeft dat deze door ieder redelijk denkend mens als privé zal worden beschouwd.'[177] Hieronder vallen foto's van een deels ontbloot lichaam waarop borsten, billen of geslachtsdelen zijn te zien of van een gekleed persoon die seksuele handelingen verricht of ondergaat.[178]

In de wetsgeschiedenis is geen aandacht besteed aan de vraag of de afbeelding echt moet zijn of ook een deepfake (of anderszins gemanipuleerd of computer-gegenereerd beeld) kan betreffen. Het moet in elk geval om een bestaande persoon gaan, die nadeel kan ondervinden van de openbaarmaking. De formulering (een afbeelding van seksuele aard van een persoon) lijkt deepfakes niet uit te sluiten: wanneer iemand een pornovideo zodanig manipuleert dat het lijkt dat een bestaande persoon daarin figureert, betreft het immers een afbeelding van seksuele aard van die persoon. Deepfake-seksvideo's zullen dan ook onder artikel 139h lid 2 onder b Sr vallen, voor zover de beelden openbaar worden gemaakt. In de meeste gevallen zal zulke openbaarmaking nadelig voor de persoon kunnen zijn, behalve als de persoon zelf toestemming heeft gegeven en de openbaarmaking bijvoorbeeld voor commercieel gewin geschiedt. De openbaarmaking hoeft overigens niet uit wraakgevoelens te geschieden: de strafbaarstelling is weliswaar gebaseerd op het fenomeen van wraakporno, maar betreft gezien de formulering en wetsgeschiedenis elke vorm van nadelige openbaarmaking van seksuele afbeeldingen, ongeacht de precieze intentie.

Naast wraakporno is in dezelfde wet ook het wederrechtelijk maken van seksuele afbeeldingen strafbaar gesteld. Artikel 139h lid 1 Sr betreft het opzettelijk en wederrechtelijk maken van een afbeelding van seksuele aard (bijvoorbeeld stiekem gemaakte foto's in douchehokjes of onder-de-rok-foto's) en het in bezit hebben van aldus gemaakte afbeeldingen. Artikel 139h lid 2 onder a Sr betreft het openbaar maken van zulke afbeeldingen, als men weet of redelijkerwijs zou moeten vermoeden dat de afbeeldingen opzettelijk en wederrechtelijk zijn gemaakt.

De vraag is nu of ook het maken van deepfakevideo's onder het wederrechtelijk maken van een afbeelding van seksuele aard kan worden verstaan. Grammaticaal is deze interpretatie mogelijk: wie een pornovideo dusdanig manipuleert dat daarin een bestaande persoon lijkt te figureren, waarbij normaliter echt beeldmateriaal van (bijvoorbeeld het hoofd van) die persoon wordt gebruikt, maakt immers een afbeelding van seksuele aard van die persoon. Volgens de Memorie van Toelichting is het maken van seksuele afbeeldingen strafbaar 'ongeacht de plaats waar dit gebeurt of het middel dat hiervoor wordt gebruikt'.[179] Ook dat zou kunnen suggereren dat het met deepfaketechnologie fabriceren van een seksfilmpje onder de strafbaarstelling valt. Toch moet deze interpretatie worden afgewezen. De voorbeelden in de wetsgeschiedenis gaan steeds uit van een fysieke situatie waarbij iemand met een foto- of videocamera beelden maakt. De ratio van de strafbaarstelling is dan ook primair het beschermen van mensen tegen het heimelijk maken van seksueel getinte foto's of video's. Belangrijk daarbij is vooral de plaatsing in het

wetboek: de wetgever heeft niet aangesloten bij de misdrijven tegen de zeden (art. 239 e.v. Sr), maar bij de misdrijven tegen de openbare orde, waarbij 'zoveel mogelijk [is] aangesloten bij de delictsomschrijving van artikel 139f Sr waarin het heimelijk filmen op een niet-openbare plaats strafbaar is gesteld'.[180] Artikel 139h lid 1 en lid 2 onder a Sr zijn dus bedoeld als aanvulling op de bestaande strafbaarstelling van heimelijke camera-opnames. Waar bij wraakporno de ratio gelegen is in het beschermen van personen tegen het in de openbaarheid rondgaan van seksuele afbeeldingen als zodanig (een ratio die ook deepfakevideo's omvat), is bij het wederrechtelijk maken van seksuele afbeeldingen de ratio gelegen in de bescherming tegen het maken van heimelijke afbeeldingen in de fysieke ruimte. Deepfakes passen niet bij die ratio, zodat hier een wetshistorische en teleologische interpretatie moet prevaleren boven de grammaticale interpretatie.

Dit betekent dat het maken van deepfakeseks video's als zodanig niet strafbaar is, maar alleen de openbaarmaking ervan (als dat nadelig voor de gedeepfakete persoon kan zijn). Of dat een lacune is, hangt af van de vraag of het voor eigen gebruik maken van deepfakeseksvideo's waarin een andere persoon figureert, strafwaardig is. Dat is een rechtspolitieke vraag, die de wetgever tot nu toe niet heeft geadresseerd. Aangezien het begrip 'openbaar maken' in artikel 139h Sr ruim is (hieronder 'valt zowel het aan één of meer personen bekend maken',[181] dus alleen al het tonen aan één andere persoon levert openbaarmaking op), betreft de eventuele lacune in de wet alleen het puur voor eigen gebruik creëren van deepfakeseksvideo's. Dat is misschien moreel verwerpelijk, maar verschilt daarin niet wezenlijk van het puur voor eigen gebruik creëren van andere typen onsmakelijke deepfakes waarin iemands afbeelding wordt gemanipuleerd, zoals een onthoofdingsvideo of een filmpje waarin iemand in een beerput valt. De rechtspolitieke vraag of het voor eigen gebruik maken van deepfakes van iemand zonder diens toestemming strafbaar zou moeten worden gesteld, betreft daarom niet zozeer seksuele deepfakes in het bijzonder, maar meer algemeen alle vormen van deepfakes die in een bepaald opzicht onterend worden geacht voor de gedeepfakete persoon.

3.1.3.5 Overige bepalingen

Naast genoemde bepalingen kunnen diverse andere artikelen van toepassing zijn. Wanneer een deepfakevideo een minderjarige (iemand onder de achttien, of die eruitziet als onder de achttien) toont met seksueel getinte inhoud, is er sprake van kinderpornografie (art. 240b Sr; zie ook art. 251 Sr conceptwetsvoorstel seksuele misdrijven).[182] Aangezien de strafbaarstelling ook virtuele kinderpornografie omvat (door het bestanddeel '*of schijnbaar* is betrokken'), zijn ook kinderpornografische deepfakevideo's strafbaar. Dat geldt ook als er geen bestaande minderjarige is afgebeeld: ook fictieve afbeeldingen zijn strafbaar, mits ze voldoende realistisch zijn om echt te lijken, wat bij deepfakes normaliter het geval zal zijn. Mutatis mutandis geldt hetzelfde voor dierenpornografie (art. 254a Sr; zie ook art. 254b Sr conceptwetsvoorstel seksuele misdrijven); aangezien ook virtuele dierenpornografie hierin strafbaar is gesteld ('een seksuele handeling waarbij een mens en een dier zijn betrokken of schijnbaar zijn betrokken', cursivering toegevoegd), zijn dierenpornografische deepfakevideo's eveneens strafbaar.
Artikel 240 onder 20 Sr (zie ook art. 151d onder b Sr conceptwetsvoorstel seksuele misdrijven)

stelt het ongevraagd toesturen van oneerbare afbeeldingen strafbaar. Dit betreft pornografische afbeeldingen die volgens de heersende seksuele moraal aanstootgevend zijn, zoals afbeeldingen van gewelddadige of anale seks.[183] Dat zal ook het ongevraagd toezenden van aanstootgevende deepfake-filmpjes omvatten: het gaat immers om een oneerbare afbeelding als zodanig, ongeacht of de afbeelding een correcte weergave van de werkelijkheid biedt. Evenzo kan ook stalking oftewel belaging (art. 285b Sr) worden gepleegd door middel van deepfakes, wanneer iemand meermaals wordt bestookt met deepfakevideo's of geluidsopnames met het doel die persoon te dwingen iets te doen of te laten of vrees aan te jagen.

Verder is het van belang te wijzen op het feit dat opzettelijke inbreuken op het portretrecht (art. 19-21 Aw, zie par. 3.4.1) volgens artikel 31 Aw strafbaar zijn, zodat het zonder toestemming openbaar maken van deepfakes die iemands portret bevatten, in beginsel strafbaar is. Daarbij moet wel worden opgemerkt dat in Nederland het auteursrecht nauwelijks strafrechtelijk wordt gehandhaafd; de privaatrechtelijke handhaving staat voorop.

3.1.4 Strafbepalingen ter bescherming van gegevens: Gegevensaantasting (art. 350a Sr)

Het derde cluster betreft strafbepalingen die primair belangen betreffende gegevens beogen te beschermen, zoals ingevoerd in de computercriminaliteitswetgeving. Het gaat om de bescherming van de vertrouwelijkheid, beschikbaarheid en **integriteit en authenticiteit van gegevens.** Bij deepfakes gaat het met name om dit laatste belang.

Bij de Wet computercriminaliteit is in 1993 gegevensaantasting strafbaar gesteld in artikel 350a lid 1 Sr:

> *'Hij die opzettelijk en wederrechtelijk gegevens die door middel van een geautomatiseerd werk of door middel van telecommunicatie zijn opgeslagen, worden verwerkt of overgedragen, verandert, wist, onbruikbaar of ontoegankelijk maakt, dan wel andere gegevens daaraan toevoegt, wordt gestraft met gevangenisstraf van ten hoogste twee jaren of geldboete van de vierde categorie.'*

Bij een deepfake waarbij iemands digitale afbeelding of stemopname wordt gebruikt, lijkt er sprake te zijn van het veranderen van gegevens. De afbeelding of het geluid wordt immers aangepast tot iets anders. De vraag is echter of dat verandering van gegevens inhoudt. Eerder lijkt bij deepfaketechnologie sprake te zijn van het gebruik van gegevens om andere gegevens te genereren; de oorspronkelijke gegevens blijven in beginsel behouden, ze worden alleen gebruikt als input voor de software om een nieuwe video of nieuw geluidsbestand te maken. Vanzelfsprekend zal het zonder toestemming vervangen van een filmpje op het YouTube-account van iemand anders door een deepfake-filmpje wel gegevensaantasting opleveren, maar het maken van een deepfake als zodanig houdt niet een aantasting van de voor de deepfake gebruikte inputgegevens in.

Wel is er bij het maken van een deepfake sprake van verandering en toevoegen van gegevens, zoals dat bij elk gebruik van een computer het geval is. De bedoeling van computergebruik (behalve fysiek gebruik als bijzettafeltje of *pressepapier*) is immers om enige verandering in de

gegevenstoestand van de computer(opslag) te bewerkstelligen. Ook het versturen of op internet plaatsen van een deepfake levert altijd verandering of toevoeging van gegevens op. Daarom is het bestanddeel 'wederrechtelijk' cruciaal in artikel 350a lid 1 Sr. Een deepfake op je eigen computer maken of op je eigen internetpagina of sociale-media-account plaatsen kan bezwaarlijk als zodanig wederrechtelijk worden genoemd.

Wel zou de handeling van het maken of verspreiden van een bepaalde deepfake een wederrechtelijke handeling kunnen opleveren – over die vraag gaat immers dit rapport. Men zou dan kunnen betogen dat die wederrechtelijke handeling (bijvoorbeeld een onrechtmatige daad of een onrechtmatige verwerking van persoonsgegevens) dan ook de handeling van het maken of verspreiden van de deepfake wederrechtelijk maakt in de zin van artikel 350a Sr. Wederrechtelijkheid kan immers worden opgevat als 'het handelen zonder daartoe gerechtigd te zijn, bijvoorbeeld zonder eigen recht of zonder toestemming van de rechthebbende'.[184] In die zin zou artikel 350a lid 1 Sr opgevat kunnen worden als een zeer breed vangnet voor alle civielrechtelijk onrechtmatige handelingen met computers die niet onder een specifieke strafbepaling vallen, bij wijze van spreken een algemene strafbaarstelling van een computer-gerelateerde onrechtmatige daad.

Die interpretatie gaat echter wel erg ver en is vermoedelijk niet zo door de wetgever bedoeld. Bij de strafbaarstelling van gegevensaantasting is toch vooral gedacht aan het zonder toestemming aantasten van gegevens in de computer of op een gegevensdrager van iemand anders. Het beschermde belang van artikel 350a Sr is primair het 'ongestoorde gebruik van gegevens', zoals de strafbaarstelling van zaakbeschadiging vooral het ongestoorde gebruik van goederen beschermt.[185] Dat laatste geldt primair voor de rechthebbende(n) op de goederen; zo zal artikel 350a Sr ook primair het belang van het ongestoorde gebruik van gegevens door de rechthebbende(n) op die gegevens betreffen. Aangezien gegevens echter, anders dan goederen, geen onderwerp van eigendomsrecht zijn, is minder duidelijk wie de rechthebbenden zijn die ongestoord gegevens moeten kunnen gebruiken. Auteursrechthebbenden zullen eronder vallen, evenals vermoedelijk verwerkers van persoonsgegevens (die immers een plicht hebben te zorgen voor de juistheid van de gegevens die zij verwerken, art. 5 lid 1 onder d AVG). Verder zal het vooral gaan om de normale gebruiker(s) van de desbetreffende gegevens. Bij het maken van een deepfake is dat de maker zelf, niet degene van wie een afbeelding of geluidsopname wordt gebruikt om deepfakes te genereren. Het maken van een deepfake belemmert in die zin niet het ongestoorde gebruik van de gegevens; het vormt integendeel juist eerder het beoogde gebruik van de gegevens. Hooguit zou artikel 350a lid 1 Sr met een creatieve interpretatie kunnen worden ingezet om civielrechtelijk onrechtmatige deepfakes strafrechtelijk aan te pakken als deze zonder toestemming gemaakt zijn op de computer van iemand anders, bijvoorbeeld op een laptop die iemand heeft uitgeleend zodat de lener aan haar scriptie kon werken nu haar eigen laptop door een ransomware-aanval ontoegankelijk is. In dat geval zijn immers gegevens op de leenlaptop veranderd of toegevoegd zonder toestemming van de rechthebbende op de laptop. Maar of zulk gebruik afbreuk doet aan het ongestoorde gebruik van de gegevens op de laptop door de uitlener is de vraag, zeker als de software en deepfakes zelf na gebruik zijn verwijderd voordat de laptop is teruggegeven.

Bij het verspreiden van deepfakes ligt de zaak misschien iets anders; men zou kunnen betogen dat bij op internet of sociale media geplaatste deepfakes degenen die regulier toegang daartoe hebben, gebruikers zijn die recht hebben op het ongestoorde gebruik van die gegevens. Zij mogen er immers tot op zekere hoogte van uitgaan dat de integriteit van de gegevens niet is aangetast. Dat betreft echter vooral de technische integriteit, dat wil zeggen dat de gepubliceerde gegevens niet in de tussentijd door onbevoegden zijn aangetast; het gaat niet om de inhoudelijke integriteit, in de zin dat de gepubliceerde gegevens zouden moeten "kloppen". Het belang van ongestoord gebruik van via internet of sociale media beschikbare gegevens kan immers bezwaarlijk een recht op "juiste" of "echte" gegevens impliceren.

Al met al zal artikel 350a lid 1 Sr, hoewel op het oog een relevante bepaling, weinig handvatten biedt om strafwaardige deepfakes te bestrijden. Van strafbare gegevensaantasting is wel sprake wanneer een deepfake zonder rechtmatige titel op iemands computer of een server wordt geplaatst, of wanneer iemand een filmpje op internet zonder toestemming vervangt door een gedeepfaket filmpje. Maar dat is niet specifiek voor deepfakes; het geldt voor alle gegevens die onrechtmatig worden geplaatst of vervangen. Op het maken of verspreiden van deepfakes als zodanig is artikel 350a Sr niet van toepassing.

Aangezien artikel 350a lid 1 Sr zelf niet of nauwelijks van toepassing is op het maken van deepfakes, is ook de strafbare voorbereidingshandeling van misbruik van hulpmiddelen (art. 350d Sr) niet van toepassing. Het maken, verspreiden of bezitten van deepfaketechnologie is dan ook niet als zodanig strafbaar.

3.1.5 Conclusie

Het 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. Verreweg de meeste financieel-economische en reputatie-gerelateerde uitingsdelicten zijn immers voldoende technologie-neutraal geformuleerd wat betreft de vorm waarin deze kunnen worden gepleegd.

Voor seksfilmpjes die met deepfaketechnologie zijn gemanipuleerd zodat het lijkt alsof een bestaande persoon daarin figureert – een veelvoorkomende en mogelijk voor slachtoffers de meest ingrijpende vorm van deepfakes – is het belangrijk te constateren dat daarop de strafbaarstelling van wraakporno (art. 139h lid 2 onder b Sr) toepasbaar is, althans voor zover zulke filmpjes openbaar worden gemaakt en dat nadelige gevolgen voor de afgebeelde persoon kan hebben. Aan die laatste voorwaarde zal meestal zijn voldaan, en aangezien openbaarmaking in deze bepaling ruim wordt uitgelegd als het delen met een of meer anderen, zal artikel 139h lid 2 onder b Sr goed kunnen worden ingezet om seksuele deepfakes tegen te gaan. Daarnaast levert het verspreiden van zulke filmpjes mogelijk ook smaadschrift op, en wanneer de video ook identificerende persoonsgegevens (zoals de naam van de gedeepfakete persoon) bevat, zal bovendien artikel 231b Sr (identiteitsfraude) van toepassing zijn.

Wanneer deepfakeseksvideo's echter niet worden verspreid, maar puur voor eigen gebruik worden gemaakt en bekeken, valt dit niet onder

een strafbepaling. Het is een rechtspolitieke vraag of het voor eigen gebruik maken van zulke deepfakes van iemand zonder diens toestemming strafbaar zou moeten worden gesteld. Daarover zou wellicht een debat kunnen worden gevoerd, dat eventueel ook breder zou kunnen worden betrokken op alle vormen van deepfakes die in een bepaald opzicht onterend kunnen worden geacht voor de gedeepfakete persoon en daarmee wellicht als "intrinsiek" strafwaardig zouden kunnen worden beschouwd. Aangezien de wetgever tot nu toe alleen het bezit van kinderpornografie en dierenpornografie als "intrinsiek" strafwaardige gegevens heeft strafbaar gesteld, en niet andere vormen van inhoud die in het maatschappelijk verkeer vaak als moreel verwerpelijk worden beschouwd (denk aan onthoofdingsvideo's of *Mein Kampf*), lijkt wel terughoudendheid gepast om het maken of bezitten van "onterende" deepfakes als zodanig strafbaar te stellen.

De enige mogelijke lacune in de wetgeving die in de analyse is geconstateerd, betreft het gat 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 als nadeel kan ontstaan. 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. Voor deepfakes levert dit niet per se een lacune in de rechtsbescherming op, aangezien zoals gezegd de meeste strafbepalingen door hun technologie-neutrale formulering van toepassing zijn. Mocht de wetgever het echter 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'. Hierdoor zou immers een algemene strafbaarstelling ontstaan van misbruik van iemands gelaat of stem als daaruit enig nadeel kan ontstaan.

3.2 Gegevensbeschermingsrecht

Naast het strafrecht is het gegevensbeschermingsrecht relevant in het kader van deepfakes. Daarbij gaat het met name om de Algemene Verordening Gegevensbescherming (AVG) van de EU en de Nederlandse Uitvoeringswet AVG (UAVG), die op bepaalde punten nadere invulling geeft aan hetgeen in de EU Verordening is bepaald. Binnen de ruimte die de AVG biedt kan Nederland zelf regelgeving aannemen met het oog op deepfakes, bijvoorbeeld door aanpassing van de UAVG. Ook kan de Autoriteit Persoonsgegevens (AP), de handhavende organisatie op het gebied van de AVG, beleidsregels en richtsnoeren uitvaardigen die door in Nederland opererende partijen in ogenschouw moeten worden genomen. Voor zover het gaat om regels uit de AVG waaraan geen invulling op Lidstatelijk niveau kan of moet worden gegeven, moeten eventuele aanpassingen aan het gegevensbeschermingskader, mocht dat wenselijk worden geacht, op EU-niveau worden gerealiseerd. Daarmee wijkt het gegevensbeschermingsrecht af van de andere drie in dit hoofdstuk bestudeerde rechtsgebieden, ten aanzien waarvan Nederland een grote

beleidsvrijheid heeft om zelf regels aan te nemen of vervolgens aan te passen.

Deze paragraaf zal geen uitputtende bespreking geven van hoe de volledige AVG en alle 99 daarin vervatte artikelen van toepassing zouden kunnen zijn op deepfaketechnologie, maar zal een aantal van de belangrijkste vragen en knelpunten adresseren. Eerst zal worden besproken in hoeverre er persoonsgegevens worden verwerkt (paragraaf 3.2.1), dan zal worden ingezoomd op de huishoudelijke exceptie (paragraaf 3.2.2). Dit zijn twee belangrijke punten die raken aan de vraag of de AVG überhaupt van toepassing is op deepfakes. Ook wordt aandacht besteed aan de vraag wie al verwerkingsverantwoordelijke kan worden gezien, dat wil zeggen, op welke partij(en) de diverse plichten uit de AVG rusten (paragraaf 3.2.3). Dan volgen paragrafen die kort stilstaan bij belangrijke beginselen van de gegevensverwerking, te weten transparantie (paragraaf 3.2.4), doel en doelbinding (paragraaf 3.2.5) en datakwaliteit (paragraaf 3.2.6). Dan volgen twee paragrafen over de legitimiteit van gegevensverwerking, ten aanzien van 'gewone' persoonsgegevens (paragraaf 3.2.7) en bijzondere of gevoelige persoonsgegevens (paragraaf 3.2.8). Tot slot wordt ingezoomd op de rechten van het datasubject en met name op het recht om vergeten te worden (paragraaf 3.2.9) en op de uitzonderingen op de AVG, met name de uitzondering die geldt in het kader van de vrijheid van meningsuiting (paragraaf 3.2.10). Tot slot volgt een korte conclusie (paragraaf 3.2.11).

3.2.1 Persoonsgegevens

Allereerst is de vraag of er met deepfakes persoonsgegevens worden verwerkt. Als dat niet het geval is, dan zijn de AVG en de UAVG überhaupt niet van toepassing. Persoonsgegevens worden in de AVG omschreven als: 'alle informatie over een geïdentificeerde of identificeerbare natuurlijke persoon („de betrokkene"); als identificeerbaar wordt beschouwd een natuurlijke persoon die direct of indirect kan worden geïdentificeerd, met name aan de hand van een identificator zoals een naam, een identificatienummer, locatiegegevens, een online identificator of van een of meer elementen die kenmerkend zijn voor de fysieke, fysiologische, genetische, psychische, economische, culturele of sociale identiteit van die natuurlijke persoon'.[186]

Daarbij is het van belang dat in principe alle data aangaande een persoon onder dit begrip vallen, of het nu gaat om gevoelige gegevens of onschuldige data, zo benadrukt de voormalige Werkgroep 29, het samenwerkingsverband van alle nationale handhavende organisaties van de EU. 'This covers of course personal information considered to be "sensitive data" in Article 8 of the directive because of its particularly risky nature, but also more general kinds of information. The term "personal data" includes information touching the individual's private and family life "stricto sensu", but also information regarding whatever types of activity is undertaken by the individual, like that concerning working relations or the economic or social behaviour of the individual. It includes therefore information on individuals, regardless of the position or capacity of those persons (as consumer, patient, employee, customer, etc).[187] Ook als een deepfake bijvoorbeeld wordt ingezet om iemand in een film iets jonger te doen laten lijken, dan worden er persoonsgegevens verwerkt.

Daarbij kan het gaan om zowel feitelijke informatie als subjectieve informatie. 'It covers

"objective" information, such as the presence of a certain substance in one's blood. It also includes "subjective" information, opinions or assessments. This latter sort of statements make up a considerable share of personal data processing in sectors such as banking, for the assessment of the reliability of borrowers ("Titius is a reliable borrower"), in insurance ("Titius is not expected to die soon") or in employment ("Titius is a good worker and merits promotion").'[188] In het kader van deepfakes is belangrijk dat het hier dus ook gaat om onwaarachtige en onjuiste informatie. De zin 'Mark Rutte is de leider van de P.v.d.A' klopt niet, maar is toch een persoonsgegeven omdat het informatie over Mark Rutte probeert over te brengen, of het beeld dat mensen van hem hebben kan beïnvloeden. Het feit dat Sigrid Kaag in een deepfake de oorlog verklaart aan China, een handeling die zij niet daadwerkelijk heeft ondernomen, betekent dus niet dat de AVG niet van toepassing is.

Ook het medium waarop de data worden bewaard of verspreid is irrelevant voor de vraag of de AVG van toepassing is. Zo benadrukt de Werkgroep 29: 'the concept of personal data includes information available in whatever form, be it alphabetical, numerical, graphical, photographical or acoustic, for example. It includes information kept on paper, as well as information stored in a computer memory by means of binary code, or on a videotape, for instance.'[189] Of het dus gaat om een deepfake die is gemaakt op basis van bestaande beelden, het manipuleren, samenvoegen of aanpassen van bestaande beeld- of audiofragmenten, of het genereren van gehele nieuwe beeld- en/of audiofragmenten is derhalve niet van belang voor de vraag of de AVG van toepassing is. Ook een virtuele avatar van Frans Timmermans zal als een verwerking van zijn persoonsgegevens hebben te gelden.

Relevant is wel of de data aan een persoon zijn te relateren. Daarbij wordt een vrij brede interpretatie gehanteerd. Ook indirecte informatie zal over het algemeen onder het begrip persoonsgegeven vallen. Dat kan bijvoorbeeld gaan over gebouwen, andere personen of evenementen waarmee een persoon kan worden geïdentificeerd of die indirect iets over hem zeggen. Als er een deepfake verschijnt waarin er binnen het Torentje vlaggen met Swastika's lijken te hangen, zal dit iets lijken te zeggen over Mark Rutte, als het huis van een burger in de fik lijkt te staan, dan zal dat een persoonsgegeven zijn over de burger in kwestie, als het kind van een echtpaar in een deepfakevideo lijkt te vertellen over zijn erbarmelijke woonsituatie, dan zal dat ook als persoonsgegeven van de ouders hebben te gelden. De Werkgroep 29 stelt dat naast het indirect aan een persoon relateren van de inhoud, het ook kan gaan om gegevens die als doel hebben over iemand te gaan of als resultaat iemand treffen. 'Despite the absence of a "content" or "purpose" element, data can be considered to "relate" to an individual because their use is likely to have an impact on a certain person's rights and interests, taking into account all the circumstances surrounding the precise case. It should be noted that it is not necessary that the potential result be a major impact. It is sufficient if the individual may be treated differently from other persons as a result of the processing of such data.'[190]

Bij verreweg de meeste deepfakes zal derhalve sprake zijn van een verwerking van persoonsgegevens. Toch zijn er vijf grensgebieden.

Ten eerste is de AVG niet van toepassing op geanonimiseerde data. Daarvan kan bijvoorbeeld sprake zijn als audio vervormd is of beelden vervaagd. 'Om te bepalen of een natuurlijke persoon identificeerbaar is, moet rekening worden gehouden met alle middelen waarvan redelijkerwijs valt te verwachten dat zij worden gebruikt door de verwerkingsverantwoordelijke of door een andere persoon om de natuurlijke persoon direct of indirect te identificeren, bijvoorbeeld selectietechnieken. Om uit te maken of van middelen redelijkerwijs valt te verwachten dat zij zullen worden gebruikt om de natuurlijke persoon te identificeren, moet rekening worden gehouden met alle objectieve factoren, zoals de kosten van en de tijd benodigd voor identificatie, met inachtneming van de beschikbare technologie op het tijdstip van verwerking en de technologische ontwikkelingen. De gegevensbeschermingsbeginselen dienen derhalve niet van toepassing te zijn op anonieme gegevens, namelijk gegevens die geen betrekking hebben op een geïdentificeerde of identificeerbare natuurlijke persoon of op persoonsgegevens die zodanig anoniem zijn gemaakt dat de betrokkene niet of niet meer identificeerbaar is. Deze verordening heeft derhalve geen betrekking op de verwerking van dergelijke anonieme gegevens, onder meer voor statistische of onderzoeksdoeleinden.'[191] Dat betekent dat als deepfakes worden gegenereerd op basis van niet-persoonlijke, maar geanonimiseerde gegevens, de AVG niet van toepassing zal zijn. Zoals besproken worden deepfakes ook ingezet om live anonieme gesprekken te voeren, waarbij een persoon bijvoorbeeld de stem en/of het gezicht van een niet bestaand persoon aanneemt. Een dergelijke toepassing zal er inderdaad in een aantal gevallen toe leiden dat de AVG niet van toepassing is. Toch is dat anders als uit bijvoorbeeld iemands woordkeuze of uit de informatie die in het gesprek wordt gedeeld toch iemands identiteit met redelijke waarschijnlijkheid kan worden achterhaald.

Vervolgens is de AVG niet van toepassing op overleden personen.'[192] Wel geeft de AVG ruimte aan lidstaten om hieromtrent regels te stellen. 'De onderhavige verordening is niet van toepassing op de persoonsgegevens van overleden personen. De lidstaten kunnen regels vaststellen betreffende de verwerking van de persoonsgegevens van overleden personen.' Daarvan heeft de Nederlandse wetgever vooralsnog echter geen gebruik gemaakt in de UAVG. Dit brengt met zich dat als een overleden persoon in een deepfake verschijnt dit niet onder de AVG zal vallen. De uitzondering is als de gegevensverwerking over de overleden persoon ook iets zegt over nog levende personen, maar dat zal bij deepfakes niet snel het geval zijn. Deze beoordeling zal per casus in de omstandigheden van het geval moeten geschieden, waarbij geldt dat des te langer een persoon overleden is, des te moeilijker het zal zijn een indirecte link aan te tonen. Waar Dalí en Juliana noch directe erfgenamen hebben, zal dat voor een fakevideo tussen Caesar en Cleopatra niet meer het geval zijn.

Ten derde is de AVG niet van toepassing op de verwerking van gegevens over rechtspersonen, zoals stichtingen, B.V.'s en verenigingen. 'De bescherming die door deze verordening wordt geboden, heeft betrekking op natuurlijke personen, ongeacht hun nationaliteit of verblijfplaats, in verband met de verwerking van hun persoonsgegevens. Deze verordening heeft geen betrekking op de verwerking van gegevens over rechtspersonen en met name als rechtspersonen gevestigde ondernemingen,

zoals de naam en de rechtsvorm van de rechtspersoon[193] en de contactgegevens van de rechtspersoon.' Toch is het zo dat als gegevens die over een rechtspersoon gaan ook iets zeggen over een natuurlijk persoon (een persoon van vlees en bloed), deze verwerking onder de AVG zal vallen. Dat is bijvoorbeeld het geval als er een deepfake verschijnt over een eenmanszaak die de naam van de eigenaar draagt, maar ook als de gegevens over de rechtspersonen direct of indirect afstralen op de leiding en/of werknemers van de rechtspersoon. Het fake NOS-journaal waarin wordt bericht dat het C.D.A te Roozendaal een dekmantel voor cocaïnesmokkel is, zal bijvoorbeeld direct iets zeggen over de personen die actief zijn in die lokale afdeling. Dit betekent dat deepfake-berichten over grote ondernemingen of landen niet snel op een natuurlijk persoon zullen afstralen, denk aan een deepfake over Shell waarin nep-olielekkages in scène worden gebracht of beelden waarin Monaco in de hens lijkt te staan, maar dat dit anders kan zijn voor deepfakes over kleinere ondernemingen, verenigingen en stichtingen. Ook dit zal van geval tot geval moeten worden beoordeeld.

Het vierde spanningsveld betreft de samenvoeging van beelden van twee of meerdere personen.[194] Alhoewel er bij het vervaardigen van dergelijke deepfakes gebruik wordt gemaakt van persoonsgegevens van meerdere mensen en hierop de AVG van toepassing zal zijn (mits aan de andere voorwaarden is voldaan) is onduidelijk of het uiteindelijke resultaat, de deepfake video die wordt verspreid, ook als een verwerking van persoonsgegevens dient te gelden en zo ja, van wiens persoonsgegevens. Als het uitgangspunt is dat de video als verwerking van persoonsgegevens van ieder van de individuen heeft te gelden wiens beeld of gelijkenis is gebruikt voor de vervaardiging van de video, dan kunnen er conflicterende belangen ontstaan. Doorslaggevend bij de beantwoording van de vraag of de uiteindelijke deepfake zelf onder de AVG valt zal zijn de mate waarin door middel van de deepfake iemand geïdentificeerd kan worden, er een bepaald beeld over iemand ontstaat of de gevolgen van de deepfake raken aan de belangen van een bestaand persoon. Bij twee gezichten die worden samengevoegd tot één kan dat onder omstandigheden het geval zijn, als het personen met duidelijke gezichtskenmerken betreft. Des te meer er gegevens van des te meer personen worden

Figuur 14-15: Using two seperate decoders & Merging of Newly Created Face

gebruikt, des te onwaarschijnlijker het wordt dat het resultaat een verwerking van persoonsgegevens zal betekenen van een of meerdere van hen (bijvoorbeeld bij de gezichten van 10 personen die worden gebruikt voor het creëren van een nieuw gezicht). Toch kan het zelfs daarbij gaan om een herkenbaar element van een van hen, bijvoorbeeld iemand met een grote moedervlek op de neus (uiteraard geldt dat de trainingsdata, de input van de verschillende foto's voor het genereren van de deepfake, wel als persoonsgegevens zijn te kwalificeren). Het punt van de samenvoeging van data van twee of meer personen zal ook lastige vragen oproepen ten aanzien van de uitvoering en implementatie van de AVG, bijvoorbeeld als persoon A recht op verwijdering inroept, maar persoon B juist geporteerd is van de deepfake.

Figuur 16-17: Niet bestaande personen gegenereerd door www.thispersondoesnotexist.com

Ten vijfde kunnen er ook persoonsgegevens werden verwerkt als het gaat om geheel gefabriceerde beelden van bijvoorbeeld Angela Merkel of om een virtuele avatar van haar. Wederom zal de mate waarmee door middel van een avatar of een nieuw gecreëerd audio- en/of beeldfragment een bestaand persoon kan worden geïdentificeerd of de verwerking van die nieuwe data een effect heeft op die bestaande persoon bepalend zijn voor de vraag of er persoonsgegevens worden verwerkt. In principe zal een persoon die helemaal fictief is, maar wel zeer realistisch is, zoals bijvoorbeeld gecreëerd door de AI van 'This Person Does not Exist', niet onder de AVG vallen (natuurlijk geldt dat als er persoonsgegevens worden gebruikt om deze gezichten te generen, de initiële verwerking wel onder de AVG zal vallen). Dat betekent onder meer dat terwijl seksuele content van virtuele kinderen wel strafbaar is onder het Wetboek van Strafrecht, dit in principe niet onder de reikwijdte van het gegevensbeschermingsrecht zal vallen.

3.2.2 Huishoudelijke exceptie

Een tweede belangrijke voorwaarde voor de toepasselijkheid van de AVG is dat geen sprake is van een van de in de AVG genoemde uitzonderingen. Van de diverse in de AVG genoemde uitzonderingen is in het kader van deepfakes in horizontale relaties het meest relevant de zogenoemde huishoudelijke exceptie. De Verordening is niet van toepassing bij 'door een natuurlijke persoon bij de uitoefening van een zuiver persoonlijke of huishoudelijke activiteit'[195] 'Deze verordening', zo verklaart een overweging uit de AVG, is niet van toepassing op de verwerking van persoonsgegevens door een natuurlijke persoon in het kader van een louter persoonlijke of huishoudelijke activiteit die als zodanig geen enkel verband houdt met een

beroeps- of handelsactiviteit. Tot persoonlijke of huishoudelijke activiteiten kunnen behoren het voeren van correspondentie of het houden van adresbestanden, het sociaal netwerk en online-activiteiten in de context van dergelijke activiteiten. Deze verordening geldt wel voor verwerkingsverantwoordelijken of verwerkers die de middelen verschaffen voor de verwerking van persoonsgegevens voor dergelijke persoonlijke of huishoudelijke activiteiten.'[196]

Het Hof van Justitie heeft twee belangrijke rechtszaken gewezen die over deze exceptie gaat. De eerste is de zaak Bodil Lindqvist uit 2003, waar een dame een soort persoonlijke hobbypagina bijhield op het internet en daar ook informatie en wetenswaardigheden over kennissen en collega's deelde, zoals onder meer dat een van hen een been had gebroken. De vraag was of een dergelijke handeling onder de huishoudelijke exceptie viel, nu het doeleinde waarvoor de gegevens werden verwerkt primair persoonlijk was, nu de internetpagina vooral voor de dame zelf en een kleine kring bekenden was bedoeld. Het Hof van Justitie ging daar echter niet in mee en stelde dat: 'Die uitzondering moet derhalve aldus worden uitgelegd, dat zij uitsluitend betrekking heeft op activiteiten die tot het persoonlijke of gezinsleven van particulieren behoren, hetgeen klaarblijkelijk niet het geval is met de verwerking van persoonsgegevens die bestaat in hun openbaarmaking op internet waardoor die gegevens voor een onbepaald aantal personen toegankelijk worden gemaakt.'[197]

Het openbaar maken van gegevens aan een onbepaalde groep mensen is in ieder geval geen verwerking voor puur persoonlijke of huishoudelijke doeleinden, al was het maar omdat de verdere verwerking niet kan worden begrensd voor wat betreft die doeleinden. Ook als persoonlijke informatie wordt gedeeld met mensen buiten een kleine kring van vrienden en familieleden zal de verwerking niet snel onder de huishoudelijke exceptie vallen. Zo gaf de Werkgroep 29 ten aanzien van Social Network Sites (SNS) aan dat die sites 'standaard en gratis privacy-vriendelijke settings dienen te hanteren die de toegang tot informatie limiteren tot de door gebruikers geselecteerde contacten. Wanneer toegang tot profielinformatie verder gaat dan deze contacten, zoals wanneer toegang tot het profiel wordt geboden aan alle deelnemers van een SNS of wanneer de data wordt geïndexeerd door zoekmachines, dan gaat de toegang verder dan de persoonlijke of huishoudelijke sfeer. Als een gebruiker zelf informatie deelt buiten de cirkel van geselecteerde vrienden, dan zal hij als verantwoordelijke worden aangemerkt. Effectief zal dan hetzelfde juridische regime van toepassing zijn als wanneer een persoon een ander technologisch platform gebruikt om persoonlijke informatie te publiceren op het web.'[198]

De tweede uitspraak betreft de zaak Ryneš uit 2013. Daarin ging het om een persoon die een camera had gericht op de toegang tot zijn erf, voor veiligheidsdoeleinden. Wederom was de vraag of deze toepassing onder de huishoudelijke exceptie viel, nu het doel van de werking van persoonsgegevens (in casu van mensen die toegang zochten tot het huis) primair van persoonlijke aard was en de gegevens niet waren bedoeld om openbaar te worden gemaakt. Toch oordeelde het Hof van Justitie ook in deze zaak anders. 'Voor zover het gebruik van een videobewakingssysteem, zoals dat in het hoofdgeding, de openbare ruimte bestrijkt – zelfs gedeeltelijk – en hierdoor buiten de privésfeer geraakt van degene die door middel van dit

systeem gegevens verwerkt, kan het niet worden beschouwd als een activiteit die met uitsluitend „persoonlijke of huishoudelijke doeleinden" wordt verricht'.[199] Daarmee lijkt er op dat het voor het Hof niet van doorslaggevend belang was wat de doeleinden van de verwerking was, maar wat de bron was van waaruit de data werden verzameld. (Overigens zou uit de Bodil Lindqvist zaak, waar het de bedoeling van de dame in kwestie leek om voor zichzelf of een hele selecte groep mensen een aantal gegevens bij te houden en zij niet doorhad of het zich niet besefte dat de webpagina voor de hele wereld toegankelijk was, kunnen worden afgeleid dat ook het doel of de bedoeling van de verwerkingsverantwoordelijke doorslaggevend is).

In het kader van deepfakes volgen uit het voorgaande vier belangrijke punten.

Allereerst is de vraag, die niet specifiek aan deepfakes is gerelateerd en meer algemeen geldt voor het gegevensbeschermingsrecht, waar precies de grens ligt tussen een besloten groep mensen en het openbaar maken van informatie. Iemand die een deepfake maakt van zijn eigen partner, van de buurvrouw of een Bekende Nederlander, maar die slechts zelf bekijkt of aan zijn huisgenoten laat zien zal vermoedelijk nog een beroep kunnen doen op deze exceptie. Al twijfelachtiger is dat waar een dergelijke deepfake in een buurtapp of groepsapp met vrienden en/of familieleden wordt gedeeld. Er is geen eenduidige grens te trekken op dit punt, maar hoe over het algemeen geldt dat des te groter en diverser de groep wordt, hoe minder snel dat de huishoudelijke exceptie van toepassing zal zijn. Daarbij moet zowel rekening worden gehouden met het feit dat er altijd een internet tussenpersoon en vaak meerdere toegang hebben tot de data, zoals de app-provider, de producent van de telefoon en de eventuele cloudprovider, als met het feit dat het steeds makkelijker wordt om gegevens uit besloten groepen voor grotere groepen mensen toegankelijk te maken.

De tweede vraag is in hoeverre uit met name de Ryneš zaak valt af te leiden dat het enkele feit dat gegevens uit de publieke ruimte wordt gehaald afdoende is om de AVG van toepassing te verklaren, zelfs als die data vervolgens slechts voor persoonlijke doeleinden worden verwerkt. Uit een strikte lezing van dit arrest lijkt een dergelijke interpretatie te volgen. Dat zou met zich brengen dat als een persoon op basis van foto's of video's die online beschikbaar zijn een deepfake maakt die hij aan niemand anders laat zien en alleen voor zijn eigen plezier bekijkt, toch aan de AVG gebonden is. Slechts als die persoon de oorspronkelijke data reeds zelf in zijn bezit had en daar een deepfake van maakt die hij zelf bekijkt of met een select gezelschap deelt zou in deze strikte lezing de huishoudelijke exceptie gelden en de AVG niet van toepassing zijn.

Ten derde is van belang dat de huishoudelijke exceptie stamt uit de Richtlijn bescherming persoonsgegevens uit 1995, die toen luidde: 'De bepalingen van deze richtlijn zijn niet van toepassing op de verwerking van persoonsgegevens [] die door een natuurlijk persoon in activiteiten met uitsluitend persoonlijke of huishoudelijke doeleinden wordt verricht.'[200] Een overweging uit de Richtlijn stelde toen dat het daarbij ging om 'activiteiten met uitsluitend persoonlijke of huishoudelijke doeleinden, bij voorbeeld correspondentie en het bijhouden van adressenbestanden'.[201] Interessant dat met name in de overweging de AVG uit 2016 een breder scala aan voorbeelden geeft ten

aanzien van mogelijke verwerkingen waarop de huishoudelijke exceptie van toepassing is, terwijl de Werkgroep 29 juist aarzelingen had bij de reikwijdte van de exceptie en toen het ontwerp van de AVG ter discussie voorlag voorstelde om deze exceptie in reikwijdte te beperken.

> *'WP 29 urges the legislature to use the process of introducing new data protection law as an opportunity to reduce as far as possible the legal uncertainty that currently surrounds various aspects of individuals' personal or household use of the internet. Access to the internet and more functional ICT has brought many positive new possibilities to individuals – for example instant access to knowledge, services and the possibility of contact with other people worldwide. However, data protection authorities are also experiencing an increasing number of complaints emanating from individuals' personal use of the internet. A typical complaint might be that a pupil has used a social networking site to say post a derogatory, inaccurate or hurtful message about a teacher. Currently some data protection authorities would reject any complaints about the pupil on the grounds that the processing of personal data involved would fall within the personal or household processing exemption. Some data protection authorities also take the view that other elements of the law – for example those relating to libel or harassment – are more appropriate instruments for dealing with issues such as 'cyber-bullying'. It is the case though that some DPAs [Data Protection Authorities] do – increasingly – take on the role of mediating individuals' internet postings.'*[202]

Eenzelfde knelpunt ontstaat bij deepfakes, zoals het voornoemde geval waarin een nare deepfake over een buurtgenoot wordt verspreid in een afgesloten buurtapp of als een deepfake pornofilm of andersoortige schadelijke deepfake in een afgesloten groep van middelbare scholieren wordt gedeeld. Noch de Europese regelgever heeft dit punt van zorg meegenomen, bijvoorbeeld door middel van een aanpassing aan de reikwijdte of formulering van de huishoudelijke exceptie in de AVG, noch de Nederlandse regelgever heeft op dit punt in de UAVG nadere regels gesteld. Dat betekent dat dergelijke verwerking van persoonsgegevens vooralsnog niet onder de reikwijdte van het gegevensverwerkingsrecht lijkt te vallen. Als dat zo is, dan zal hiervoor toevlucht moeten worden gezocht tot het strafrecht, het portretrecht, het recht op de bescherming van de eer en goede naam en het aansprakelijkheidsrecht.

Ten vierde en tot slot is een van de opties die de Werkgroep 29 reeds in 2013 meegaf aan de EU-wetgever – en toen overigens ook met redenen omkleed als een minder favoriete optie naar voren schoof dan het behouden van de huishoudelijke exceptie, maar die beter omschrijven en inkaderen – was het geheel of gedeeltelijk afschaffen van de huishoudelijke exceptie. 'The regulation should differ from the current Directive in that all processing of personal data performed – even for exclusively personal or household purposes – should to some extent come within the scope of the Regulation.'[203] Deze optie vond de toenmalige Werkgroep het overwegen waard omdat in 2013 de situatie fundamenteel was veranderd sinds 1995. Een van de belangrijkste veranderingen was niet alleen dat bijna iedereen toegang had tot digitale verwerkingstechnieken, maar ook het gemak waarmee data vanuit de huishoudelijke sfeer openbaar kunnen worden gemaakt. Deze tendens heeft zich sindsdien nog sterker voortgezet.

Het feit dat de EU-wetgever daar toentertijd niet voor heeft gekozen brengt met zich dat veel deepfakes vrijelijk kunnen worden gemaakt (want doorgaans in de huishoudelijke sfeer), in ieder geval voor zover gebruik wordt gemaakt van foto's of filmpjes die in de privésfeer zijn vervaardigd, en pas kunnen worden geadresseerd onder het gegevensbeschermingsrecht als die reeds in grotere kring worden verspreid. Het is waar, zoals de Werkgroep 29 toen al opmerkte, dat het afschaffen van de huishoudelijke exceptie een extra last zou leggen op de handhavende organisaties, omdat simpelweg alle verwerkingen van persoonsgegevens onder hun verantwoordelijkheid zou komen te vallen. Anderzijds is nu een situatie ontstaan waarin eventuele latere problemen niet bij de bron kunnen worden aangepakt, de handhavende organisaties steeds achter de spreekwoordelijke feiten aan lopen en er een veelheid van mogelijk onrechtmatig materiaal kan worden gecreëerd.

3.2.3 Gegevensverwerkings-verantwoordelijke

Als de gegevens die worden verwerkt juridisch gezien als 'persoonsgegeven' zijn te kwalificeren en de huishoudelijke exceptie niet van toepassing is dan is de AVG van toepassing, mits de verwerkingsverantwoordelijke, dat wil zeggen degene die het doel en de middelen bepaalt voor de gegevensverwerking, of de verwerker, de eventuele persoon of organisatie die door de verwerkingsverantwoordelijke wordt gevraagd om een deel van de verwerking voor zijn rekening te nemen, op EU-grondgebied gevestigd is. Als dat niet het geval is dan is de AVG in principe niet van toepassing. Dat betekent dat als een burger uit bijvoorbeeld de Verenigde Staten, Rusland of Marokko een deepfake maakt van een Nederlandse burger, de AVG in principe niet van toepassing zal zijn.[204]

Het is belang om kort stil te staan bij het onderscheid in de AVG tussen de posities van de gegevensverwerkingsverantwoordelijke, de verwerker en het datasubject. De terminologie is een beetje ongelukkig gekozen – zowel de verwerker als de verantwoordelijke 'verwerken' namelijk persoonsgegevens. Een verwerker verwerkt de gegevens in opdracht van de verantwoordelijke, maar de verantwoordelijke verwerkt de gegevens ook. Het onderscheid tussen de twee posities is belangrijk omdat op de verwerker minder plichten rusten dan op de verantwoordelijke. Vaak zal het redelijk evident zijn wie bij de vervaardiging en/of verspreiding van een deepfake als verwerker en wie als verantwoordelijke gezien moet worden; toch zullen er ook grensgevallen zijn die voor meerderlei interpretatie vatbaar zijn.

De verwerker is degene die in opdracht van een ander gegevens verwerkt, bijvoorbeeld op basis van een contractuele overeenkomst. Bijvoorbeeld: bedrijf A krijgt de persoonsgegevens direct van organisatie B of organisatie B bepaalt nauwgezet hoe en met welke middelen bedrijf A gegevens moet vastleggen. In dat geval is A een 'verwerker'. Hoe zelfstandiger A in dit opzicht is, dus hoe meer vrijheid hij heeft om meer gegevens te verwerken of te bepalen hoe het verwerkingsproces moet worden ingericht, hoe eerder A niet als 'verwerker' zal worden gezien, maar als 'verantwoordelijke'. Dat is een graduele lijn, waar geen exacte criteria voor zijn te geven. Factoren die hierbij van belang zijn:

Is er een contract of concrete afspraak tussen bedrijf A en bedrijf B waarin is vervat dat bedrijf A voor bedrijf B persoonsgegevens verwerkt? Zo ja, dan zal bedrijf A vaak zijn aan te merken als verwerker.

Als bedrijf A zelf een belang heeft bij het verwerken van de gegevens (bijvoorbeeld, als het ook een analyse doet op de data om profielen te maken), dan zal bedrijf A doorgaans, in ieder geval voor de verwerking van persoonsgegevens voor dat doel, als 'verantwoordelijke' worden aangemerkt.

Als bedrijf A keuzevrijheid heeft in hoe de gegevens worden verzameld en op welke wijze ze worden verwerkt (bijvoorbeeld ten aanzien van het programma dat wordt gebruikt voor de verwerking, waar de gegevens worden verzameld en hoe, in welke categorieën de gegevens worden geplaatst, etc.) dan kan het zijn dat het als verantwoordelijke zal worden gezien.

Om vast te stellen of iemand een 'verantwoordelijke' is gelden globaal twee criteria:[205]

- 1. De verantwoordelijke is degene die het doel van de gegevensverwerking bepaalt: bijvoorbeeld persoonsgegevens worden gebruikt voor het mailen van klanten, voor het verbeteren van de website of voor het ontwikkelen van nieuwe producten.
- 2. De verantwoordelijke is degene die de middelen vaststelt: dit criterium heeft betrekking op hoe de gegevens worden verzameld en verwerkt, met welke methoden en met behulp van welke technieken en software ze worden geanalyseerd.

De verantwoordelijke moet aan alle plichten uit de Verordening voldoen. Vaak zullen organisaties die samenwerken bij het verwerken van persoonsgegevens samen de verantwoordelijkheid dragen. Als bedrijf B samen met een andere organisatie, bedrijf C, de doelen en de middelen vaststelt, of bedrijf B het doel vaststelt en bedrijf C gaat over hoe en op welke wijze de persoonsgegevens worden verwerkt, dan zijn bedrijven B en C doorgaans beide als 'verantwoordelijke' aan te merken. Dat betekent dat zij een gezamenlijke verantwoordelijkheid hebben om aan alle bepalingen uit de Verordening te voldoen. Hoe die verantwoordelijkheidsverdeling precies uitvalt hangt van de situatie af. Als bedrijf B alleen zeggenschap en controle heeft over een klein gedeelte van het verwerkingsproces, dan zal het voor dat gedeelte als verantwoordelijke aan te merken zijn.

Als het hoofdkantoor Z is gevestigd in Amerika of een ander land buiten de EU en dochterorganisatie Y staat in Nederland, dan hangt het van de situatie af wie er als verantwoordelijke valt aan te merken. Daarbij is wederom de beleidsvrijheid die organisatie Y heeft bepalend. Als die zelf de ruimte heeft om te kiezen of en in hoeverre er persoonsgegevens worden verzameld en hoe, dan is organisatie Y in principe zelf de verantwoordelijke. Slechts als Y echt in opdracht handelt van het hoofdkantoor en geen zeggenschap heeft over welke gegevens er worden verzameld, waarom en hoe, kan dat anders zijn.

Tot slot is er dan ook de positie van het datasubject, ook wel de betrokkene. Dat is degene over wie persoonsgegevens worden verwerkt. Het datasubject wordt beschermd in de AVG en heeft tal van rechten, zoals het recht op gegevenswissing, rectificatie en informatie over de doeleinden en wijzen van verwerking.

Alhoewel deze posities in theorie goed uit elkaar te halen zijn lopen die in de praktijk vaak door elkaar, zo ook bij deepfakes. Als één partij de technologie levert voor deepfakes, de andere een App op de markt brengt, een derde het in een App-store plaatst, een vierde zorgt voor de cloud toepassing om gegevens op te slaan, een vijfde een platform of sociaal medium waarop de deepfake wordt verspreid, etc., wie is dan precies de verwerker en wie de verantwoordelijke, of als er meerdere verantwoordelijken zijn, op welke verantwoordelijken rusten dan precies welke AVG-plichten? Daarbij komt dat de burger vaak net zo goed persoonsgegevens van anderen verwerkt, bijvoorbeeld als hij een deepfake van een vriend of van een Bekende Nederlander vervaardigd. Op een deepfake waarin zowel hij als een hem bekend persoon zijn afgebeeld. In zulke gevallen zal hij vermoedelijk als gedeelde verwerkingsverantwoordelijke worden aangemerkt, omdat hij in ieder geval het doel van de verwerking in een specifiek geval bepaalt. Ook komt het voor dat hij een deepfake alleen van zichzelf maakt. Daarmee is hij strikt genomen ook een verwerkingsverantwoordelijke, maar zoals gezegd zal daar in de praktijk weinig gevolg aan worden gegeven.

3.2.4 Transparantie

Als het EU-gegevensbeschermingsrecht van toepassing is dan vereist de AVG dat de gegevensverwerking transparant geschiedt.[206] Dat betekent niet alleen dat het data subject – degene waarover de persoonsgegevens gaan – het recht heeft om informatie te vragen aan de verwerkingsverantwoordelijke over het hoe, waarom en wat van de gegevensverwerking,[207] maar ook dat de verwerkingsverantwoordelijke een directe plicht heeft om het datasubject actief en op eigen initiatief dergelijke informatie te verstrekken. Daarbij onderscheidt de AVG twee situaties. Ten eerste het geval waarin de data van het datasubject zijn verkregen, dat wil zeggen dat de verantwoordelijke in direct contact stond met het datasubject. Hiervan zal bijvoorbeeld sprake zijn als een persoon een foto heeft gemaakt van zijn geliefde of die een selfie heeft gemaakt en die heeft toegestuurd. Ten tweede het geval waarin de data 'niet van de betrokkene zijn verkregen', waar bijvoorbeeld sprake van is in het geval data van het internet worden gehaald.

In het eerste geval moet de verwerkingsverantwoordelijke op het moment dat hij die gegevens verzamelt het datasubject ervan op de hoogte stellen waarom hij dat doet en voor welke doeleinden hij dat doet. Hier wreekt zich gelijk het punt van de huishoudelijke exceptie. De meeste foto's, films of geluidsfragmenten die in de privésfeer worden verzameld zullen onder de huidige interpretatie van het gegevensbeschermingsrecht vermoedelijk onder deze exceptie vallen. Op dat moment is de AVG dus niet van toepassing en is de persoon die de foto maakt of verkrijgt niet aan de informatieplicht gebonden. Op het moment dat hij er een deepfake van maakt en deze verspreidt is de AVG echter wel van toepassing. Dan zou in het geval hij de data oorspronkelijk van het datasubject heeft verkregen, de informatieplicht vanaf dat moment gelden. Dat betekent dat de verwerkingsverantwoordelijke een substantiële hoeveelheid informatie moet verstrekken aan het datasubject. Omdat deze bepaling uitgaat van direct contact tussen de verwerkingsverantwoordelijke en het datasubject, waarbij de AVG direct van toepassing is, is niet in een uitzondering op de informatieplicht voorzien. Dat betekent dat als persoon A van persoon B een foto heeft gemaakt en vervolgens persoon

B uit het oog is verloren, hij niet mag overgaan tot het verspreiden van een eventuele deepfake op basis van die foto voordat hij persoon B heeft gecontacteerd en alle relevante informatie heeft verstrekt.[208]

> *'1.Wanneer persoonsgegevens betreffende een betrokkene bij die persoon worden verzameld, verstrekt de verwerkingsverantwoordelijke de betrokkene bij de verkrijging van de persoonsgegevens al de volgende informatie:*
> *de identiteit en de contactgegevens van de verwerkingsverantwoordelijke en, in voorkomend geval, van de vertegenwoordiger van de verwerkingsverantwoordelijke; []*
> *de verwerkingsdoeleinden waarvoor de persoonsgegevens zijn bestemd, alsook de rechtsgrond voor de verwerking;*
> *de gerechtvaardigde belangen van de verwerkingsverantwoordelijke of van een derde, indien de verwerking op artikel 6, lid 1, punt f), is gebaseerd;*
> *in voorkomend geval, de ontvangers of categorieën van ontvangers van de persoonsgegevens; []*
>
> *2. Naast de in lid 1 bedoelde informatie verstrekt de verwerkingsverantwoordelijke de betrokkene bij de verkrijging van de persoonsgegevens de volgende aanvullende informatie om een behoorlijke en transparante verwerking te waarborgen:*
> *a) de periode gedurende welke de persoonsgegevens zullen worden opgeslagen, of indien dat niet mogelijk is, de criteria ter bepaling van die termijn;*
> *b) dat de betrokkene het recht heeft de verwerkingsverantwoordelijke te verzoeken om inzage van en rectificatie of wissing van de persoonsgegevens of beperking van de hem betreffende verwerking, alsmede het recht tegen de verwerking bezwaar te maken en het recht op gegevensoverdraagbaarheid;*
> *c) wanneer de verwerking op artikel 6, lid 1, punt a), of artikel 9, lid 2, punt a), is gebaseerd, dat de betrokkene het recht heeft de toestemming te allen tijde in te trekken, zonder dat dit afbreuk doet aan de rechtmatigheid van de verwerking op basis van de toestemming vóór de intrekking daarvan;*
> *d) dat de betrokkene het recht heeft klacht in te dienen bij een toezichthoudende autoriteit; []'*[209]

Deze plicht om vooraf de relevante informatie te verstrekken blijkt des te meer uit een additioneel punt, dat raakt aan het in de volgende sub–paragraaf genoemde principe van doel en doelbinding, namelijk dat als de data voor een ander doel worden verwerkt dan oorspronkelijk het geval was, wat vaak het geval zal zijn bij deepfakes omdat datasubjecten zelden informatie over zichzelf zullen verstrekken met het doel deze tot deepfake te verwerken, er een extra plicht geldt. 'Wanneer de verwerkingsverantwoordelijke voornemens is de persoonsgegevens verder te verwerken voor een ander doel dan dat waarvoor de persoonsgegevens zijn verzameld, verstrekt de verwerkingsverantwoordelijke de betrokkene vóór die verdere verwerking informatie over dat andere doel en alle relevante verdere informatie als bedoeld in lid 2.'[210]

Ten tweede is er het geval waarin de data niet van het datasubject zijn verkregen. Deze bepaling geldt niet alleen als de gegevens uit een openbare bron zijn verkregen, maar ook als de data indirect van het datasubject zijn verkregen, zoals wanneer persoon A aan persoon B vraagt om een foto van persoon C, de partner van B, toe te sturen. In dergelijke gevallen geldt dezelfde

moeilijkheid van de huishoudelijke exceptie. Zoals hierboven benoemd geldt de plicht om grosso modo dezelfde informatie te verstrekken als bovenstaand geciteerd, met als belangrijk additioneel punt dat ook moet worden vermeld hoe de data zijn verkregen, en geldt dezelfde bepaling als gegevens worden verwerkt voor een ander doel. Wel gelden er twee afwijkende punten ten aanzien van het geval waarin de data van het datasubject zelf afkomstig zijn.

Enerzijds betreft dat de termijn. Daarover stelt de AVG 'De verwerkingsverantwoordelijke verstrekt de in de leden 1 en 2 bedoelde informatie: a) binnen een redelijke termijn, maar uiterlijk binnen één maand na de verkrijging van de persoonsgegevens, afhankelijk van de concrete omstandigheden waarin de persoonsgegevens worden verwerkt; b) indien de persoonsgegevens zullen worden gebruikt voor communicatie met de betrokkene, uiterlijk op het moment van het eerste contact met de betrokkene; of c) indien verstrekking van de gegevens aan een andere ontvanger wordt overwogen, uiterlijk op het tijdstip waarop de persoonsgegevens voor het eerst worden verstrekt.'[211] Bij 'ontvanger' moet primair worden gedacht aan een persoon of organisatie waaraan de data direct worden verspreid, zoals het geval is als een deepfake wordt doorgestuurd aan bekenden.[212] Toch zal deze grond vermoedelijk ook van toepassing worden verklaard in het geval de deepfake op het internet wordt gepubliceerd, omdat anders de situatie zou ontstaan dat een verwerkingsverantwoordelijke aan striktere regels dient te voldoen (het verstrekken van de informatie aan het data subject dus het moment dat de deepfake aan anderen wordt toegestuurd) dan als hij het op het internet publiceert (het verstrekken van de informatie uiterlijk na een maand). Hoe dan ook geldt dat hierbij uiterlijk na een maand de relevante informatie aan het datasubject zal moeten worden verstrekt.

Anderzijds geldt op de informatieplicht een belangrijke uitzondering. Deze plicht geldt niet wanneer en in zoverre 'het verstrekken van die informatie onmogelijk blijkt of onevenredig veel inspanning zou vergen, in het bijzonder bij verwerking met het oog op archivering in het algemeen belang, wetenschappelijk of historisch onderzoek of statistische doeleinden, behoudens de in artikel 89, lid 1, bedoelde voorwaarden en waarborgen, of voor zover de in lid 1 van dit artikel bedoelde verplichting de verwezenlijking van de doeleinden van die verwerking onmogelijk dreigt te maken of ernstig in het gedrang dreigt te brengen. In dergelijke gevallen neemt de verwerkingsverantwoordelijke passende maatregelen om de rechten, de vrijheden en de gerechtvaardigde belangen van de betrokkene te beschermen, waaronder het openbaar maken van de informatie.'[213] Duidelijk is dat bij deze uitzondering primair wordt gedacht aan het geval dat er een grote dataset wordt verwerkt voor een publiek belang, zoals het geval waarin onderzoekers van een universiteit toegang krijgen tot de klantgegevens van de V&D om studie te doen naar de mogelijke oorzaken van het faillissement. De data kunnen in zo'n geval zoveel personen betreffen dat ieder van hen individueel contacteren vrijwel ondoenlijk is, of in ieder geval zoveel inspanning zou vergen dat de onderzoekers dan vermoedelijk zouden afzien van hun studie. Daarbij moet onder andere in ogenschouw worden genomen dat de eventuele contactgegevens die in de bestanden zijn opgenomen in veel gevallen reeds verouderd zullen zijn. In zo'n geval kan een uitzondering worden gemaakt op de informatieplicht.

Zelfs in het geval een rechter zou oordelen dat deze uitzondering niet slechts opgaat voor verwerkingen in het kader van een algemeen belang, zoals een archiveringstaak of wetenschappelijk of statistisch onderzoek, dan nog geldt dat het bij deepfakes vrijwel altijd om een of enkele personen gaat, waardoor het onwaarschijnlijker is dat het hier zal gaan om 'onevenredige inspanningen', alhoewel een rechter dat mogelijk zal relateren aan de ernst van de deepfake. Als het om een deepfake gaat die iemands reputatie schaadt of anderszins schade berokkent zal dit niet snel het geval zijn, maar het is niet uitgesloten dat een rechter deze uitzondering van toepassing zal verklaren op een onschadelijke en grappige deepfake. Daarbij moet evenwel worden bedacht dat als het om een Bekende Nederlander gaat, het vrijwel altijd mogelijk is om direct of indirect (bijvoorbeeld via management, uitgever of andere vertegenwoordigende organisatie) contact op te nemen met die persoon en dit een zeer geringe inspanning vergt; ook bij direct bekenden, vrienden of familieleden is dat het geval. Zo bezien zal deze uitzondering slechts in uitzonderlijke gevallen opgaan in het kader van deepfakes in horizontale relaties. Zelfs als deze uitzondering echter opgaat, dan stelt de AVG dat er alsnog op een andere wijze aan de informatieplicht moet worden voldaan, namelijk door de relevante informatie openbaar te maken. De verwerkingsverantwoordelijke zal dan bijvoorbeeld op een openbare website alle bovengenoemde informatie moeten verstrekken.

3.2.5 Doel en doelbinding

Vervolgens benadrukt de AVG de principes van doel en doelbinding. Persoonsgegevens moeten 'voor welbepaalde, uitdrukkelijk omschreven en gerechtvaardigde doeleinden worden verzameld en mogen vervolgens niet verder op een met die doeleinden onverenigbare wijze worden verwerkt'.[214] Dit doel moet helder en afgebakend zijn, omdat daaraan ook de principes van dataminimalisatie en opslagbeperking zijn verbonden. Het kan hierbij dus niet gaan om hele brede beschrijvingen van doeleinden. 'For these reasons, a purpose that is vague or general, such as for instance 'improving users' experience', 'marketing purposes', 'IT-security purposes' or 'future research' will - without more detail - usually not meet the criteria of being 'specific'.'[215]

In verreweg de meeste gevallen zullen er data voor het vervaardigen van deepfakes worden gebruikt die aanvankelijk voor een ander doel waren bestemd. Foto's, video's en geluidsfragmenten die in de privésfeer worden opgenomen zullen doorgaans niet met de expliciete bedoeling worden verzameld om daar later deepfakes van te maken. Dit geldt des te meer voor foto's, video's en geluidsfragmenten die zijn verzameld uit openbare bron, zoals het internet. Daarbij geldt de extra belemmering dat niet eenvoudig achterhaald kan worden wat de oorspronkelijke bedoeling van de verwerking was, terwijl de plicht bij de verwerkingsverantwoordelijke ligt om te verifiëren dat hij gegevens niet voor andere doeleinden gebruikt.

Daarbij moeten evenwel twee punten worden opgemerkt.

Enerzijds wordt hier gesproken van een 'onverenigbaar' nieuw doel. Wat dat precies betekent en in welke gevallen een nieuw doel verenigbaar is met het oorspronkelijke doel of niet, daarover bestaat een levendig debat. De Werkgroep 29 geeft als voorbeeld van duidelijk ´verenigbare´ verwerking: ´ A customer

contracts an online retailer to deliver an organic vegetable box each week to their home. After initial 'collection' of the customer's address and banking information, these data are 'further processed' by the retailer each week for payment and delivery. This obviously complies with the principle of purpose limitation and requires no further analysis.´ Als voorbeeld van een grensgeval, waarvoor nader onderzoek nodig is: 'The vegetable box retailer wishes to use the customer's email address and purchase history to send them personalized offers and discount vouchers for similar products including its range of organic dairy products. He also wishes to provide the customer's data including their name, email address, phone number, and purchase history to a business contact which has opened an organic butchery business in the neighbourhood. In both cases, the retailer cannot assume that this further use is compatible and some additional analysis is necessary, with the possibility of different outcomes (e.g. in case of 'internal' use or transfer of data.' En als voorbeeld van duidelijk onverenigbare verdere verwerking. 'The vegetable box customer also buys a range of other organic products on the retailer's website, some of which are discounted. The retailer, without informing the customer, has implemented an off-the-shelf price-customization software solution, which - among other things - detects whether the customer is using an Apple computer or a Windows PC. The retailer then automatically gives greater discounts to Windows users. In this case, the further use of available data and the unfair collection of additional information, both for an unrelated purpose (allowing secret 'price discrimination'), are problematic.'[216]

Als deze interpretatie wordt gevolgd dat zullen in veel gevallen deepfakes onverenigbaar zijn met het oorspronkelijke doel waarvoor de data werken verwerkt. Het maken van een foto als mooie herinnering aan een reis, een feest of een wandeling, zoals in de privésfeer vaak het geval zal zijn, is immers iets wezenlijk anders dan daar een deepfake van genereren. Toch is dat zeker ook niet uitgesloten. Als iemand op basis van bestaande foto's van een reis die hij heeft ondernomen met zijn geliefde een deepfakevideo maakt, waarin het lijkt alsof zij weer lopen door de straten van Venetië, dan kan dit in lijn zijn met de oorspronkelijke bedoeling van de verwerking van de foto's. Bij data die op het internet beschikbaar zijn zal het vinden van een dergelijke verenigbare verwerking lastiger voor te stellen zijn, maar ook niet uitgesloten. Als een parlementariër bijvoorbeeld een filmpje van zichzelf plaatst op zijn persoonlijke website waarin hij te zien is als hij kritische vragen stelt aan een minister en een burger monteert het filmpje zo dat het lijkt alsof Gandalf steeds achter de parlementariër steeds en hem magische krachten toestuurt, dan is zo'n bewerking niet per definitie in strijd met het oorspronkelijke doel.

Anderzijds kan, in het geval het doel niet verenigbaar is met het oorspronkelijke doel van verwerking, een nieuwe verwerkingsgrond worden gevonden in toestemming van het datasubject voor het maken van de deepfake of in een publiek belang waarvoor de deepfake nodig zou zijn. Dat kan bijvoorbeeld simpelweg het geval zijn als de persoon A aan persoon B vraagt of het ok is als hij de eerder genomen vakantie gebruikt voor een deepfake filmpje of als de politie op basis van bestaande beelden een deepfake maakt van een plaats delict. In dat geval is er simpelweg een nieuw doel voor de verwerking van persoonsgegevens gevonden. In hoeverre het vinden van een nieuw verwerkingsdoel met

een daarbij horende verwerkingsgrondslag echter is toegestaan op basis van andere verwerkingsdoelen dan toestemming en/of publiek belang blijkt niet duidelijk uit de AVG. Dit zal nader worden besproken in paragraaf 3.2.6.

3.2.6 Datakwaliteit

De AVG legt nadruk op het verwerken van gegevens die juist en volledig zijn. Zo heeft het datasubject het recht 'om van de verwerkingsverantwoordelijke onverwijld rectificatie van hem betreffende onjuiste persoonsgegevens te verkrijgen. Met inachtneming van de doeleinden van de verwerking heeft de betrokkene het recht vervollediging van onvolledige persoonsgegevens te verkrijgen, onder meer door een aanvullende verklaring te verstrekken.'[217] Als data incorrect zijn, mag het datasubject vragen ze aan te passen. Het kan hier bijvoorbeeld gaan om zijn leeftijd, of een adres dat verkeerd is ingevoerd in het systeem of verouderd is geraakt. Het recht om incorrecte gegevens te rectificeren moet in principe altijd worden gehonoreerd, tenzij het datasubject niet kan worden geïdentificeerd, of dat het datasubject om de week vraagt om een klein of betekenisloos feit te corrigeren.[218]

Omdat deepfakes per definitie nep zijn lijkt het aannemelijk dat datasubjecten immer het recht hebben om het beeld weer recht te zetten, maar zeker is dat niet. Bovendien is de vraag wie uiteindelijk het laatste woord heeft over wat correcte gegevens zijn en wat niet. Datasubjecten kunnen immers in sommige gevallen belang hebben bij het verstrekken van incorrecte gegevens, bijvoorbeeld om gunstig geprofileerd te worden door een bank, of om in bepaalde categorieën te vallen, of omdat zij zich schamen voor het feit dat zij een medische aandoening hebben. De AVG zwijgt over de vraag waar de bewijslast ligt voor het aantonen dat persoonsgegevens kloppen of niet en welke bewijsstandaard daarbij dient te worden gehanteerd. Als hierover onenigheid blijft bestaan zal dit uiteindelijk aan de AP of de rechter moeten voorgelegd. Eenzelfde punt speelt bij het aanleveren van additionele gegevens, wat bij deepfakes vermoedelijk een minder grote rol van betekenis zal spelen.

Als de verantwoordelijke een verzoek om rectificatie of aanvulling van gegevens honoreert geldt nog een additionele plicht: 'De verwerkingsverantwoordelijke stelt iedere ontvanger aan wie persoonsgegevens zijn verstrekt, in kennis van elke rectificatie of wissing van persoonsgegevens of beperking van de verwerking [], tenzij dit onmogelijk blijkt of onevenredig veel inspanning vergt. De verwerkingsverantwoordelijke verstrekt de betrokkene informatie over deze ontvangers indien de betrokkene hierom verzoekt.'[219] Stel dus dat een verantwoordelijke een verzoek krijgt tot rectificatie van de deepfake, namelijk om het 'fake' element eruit te halen, dan zal hij dat in vrijwel alle gevallen moeten honoreren. Heeft hij de deepfake verspreid, dan zijn er twee situaties. Hij weet aan wie hij de deepfake heeft doorgestuurd of kan eenvoudig achterhalen wie de deepfake nu in bezit hebben. In dat geval moet hij hen ervan op de hoogte stellen dat ook zij de deepfakevideo zullen moeten aanpassen. Dat geldt ook als hij de deepfake openbaar heeft gemaakt, maar eenvoudig kan achterhalen wie een kopie van het nepbericht heeft gemaakt. Slechts als dit niet eenvoudig te achterhalen is en er bijvoorbeeld meerdere kopieën van de deepfake zijn gemaakt kan hij van deze informatieplicht worden ontheven. Dat

zal sneller het geval zijn bij relatief onschuldige deepfakes dan bij nepberichten die bijvoorbeeld pornografische content bevatten; het gaat immers om de 'evenredigheid' van de inspanning. De evenredigheid is onder meer gerelateerd aan de schade die het datasubject ondervindt van het voortbestaan van de incorrecte kopieën.

Wellicht nog belangrijker is dat een verantwoordelijke niet slechts de plicht heeft om gegevens te rectificeren op verzoek van het datasubject, maar dat hij een zelfstandige plicht heeft om te zorgen voor de kwaliteit van de gegevens die hij verwerkt. De persoonsgegevens moeten 'juist zijn en zo nodig worden geactualiseerd; alle redelijke maatregelen moeten worden genomen om de persoonsgegevens die, gelet op de doeleinden waarvoor zij worden verwerkt, onjuist zijn, onverwijld te wissen of te rectificeren („juistheid"); [] De verwerkingsverantwoordelijke is verantwoordelijk voor de naleving van [deze plicht] en kan deze aantonen („verantwoordingsplicht").'[220] Een strikte lezing van deze plicht zou met zich brengen dat alle deepfakes per definitie verboden zijn aangezien de verantwoordelijke weet dat hij onjuiste gegevens verwerkt (natuurlijk mits de AVG van toepassing is, zie paraaf 3.2.1 en 3.2.2). Het gaat hier immers niet om juistheid in relatie tot het doel (het vervaardigen van een deepfake), maar om de juistheid als zodanig. Ook spreekt de AVG niet van een mogelijkheid voor het datasubject om de verwerkingsverantwoordelijke te ontheffen van deze plicht, wat deels te maken kan hebben met het feit dat het gegevensbeschermingsrecht niet slechts de belangen van het datasubject dient, maar ook algemene belangen die bijvoorbeeld gerelateerd zijn aan het verwerken van gegevens die waarachtig zijn. Stel Mark Rutte gaat akkoord met het maken van een fakefilm van hem, waarin het lijkt alsof hij tegen mevrouw Jorritsma en mevrouw Ollongren de loftrompet steekt over Pieter Omtzigt. Dan is het alsnog de vraag of hier het datakwaliteitsbeginsel niet wordt geschonden. Gaat het hier om een onjuiste en onvolledige voorstelling van zaken, of gaat het hier om een hypothetische en fictieve casus, die niet binnen het bereik van het datakwaliteitsbeginsel valt?

3.2.7 Legitieme verwerking persoonsgegevens

De AVG stelt dat de verantwoordelijke moet zorgdragen voor een legitieme verwerkingsgrond. Daarbij somt de Verordening exhaustief zes mogelijke gronden op. In het kader van deepfakes zullen daarvan drie relevant zijn: toestemming voor de dataverwerking door het datasubject, een contractuele overeenstemming tussen het datasubject en de verwerkingsverantwoordelijke waarvoor de verwerking van persoonsgegevens noodzakelijk is of een gerechtvaardigd belang van de verwerkingsverantwoordelijke, dat het belang van het datasubject om zijn data niet verwerkt te zien overstijgt.

Twee van de overige drie gronden betreffen gegevensverwerking in het publiek belang of als voorgeschreven bij wet. Het gaat daarbij in principe om overheidsorganisaties die in het kader van een publiek belang persoonsgegevens over burgers verzamelen. Denk daarbij bijvoorbeeld aan de Belastingdienst. In een aantal gevallen voeren ook private organisaties publieke taken uit en kunnen ook zij op deze gronden een beroep doen. Denk daarbij bijvoorbeeld aan de Nederlandse Spoorwegen. Voor burgers zal dat echter vrijwel nimmer het geval zijn. Omdat dit onderzoek zich primair richt op horizontale

relaties zullen deze verwerkingsgronden buiten beschouwing blijven. De laatste grond die vrijwel nimmer relevant zal zijn bij deepfakes in horizontale verhoudingen betreft de bescherming van het vitale belang van het datasubject, waarbij geldt dat die niet in staat is om op dat moment zijn toestemming te geven. Het kan hierbij bijvoorbeeld gaan om een medeburger die de persoonsgegevens van een pas flauwgevallen persoon doorgeeft aan de ambulance. Niet direct in te zien valt hoe deepfakes in dergelijke context acuut en noodzakelijk zullen zijn. Daarom wordt in het hiernavolgende slechts stilgestaan bij toestemming, contractuele overeenstemming en het gerechtvaardigd belang als mogelijke legitieme verwerkingsgrond.

Twee van de drie relevante gronden in het kader van deepfakes in horizontale verhoudingen zijn gebaseerd op de instemming van het datasubject, ofwel door mondelinge of schriftelijke toestemming ofwel als onderdeel van een contractuele relatie. Daarbij moet het gaan om toestemming voor de gegevensverwerking ten behoeve van het maken van deepfakes of om toestemming waaruit akkoord met voor het maken van deepfakes logischerwijs volgt. Op deze gronden kan een beroep worden gedaan als de verwerkingsverantwoordelijke – degene die de deepfake maakt en/of verspreidt – en het datasubject – degene waarover de deepfake gaat – elkaar kennen en het datasubject graag wil dat er een deepfake van hem wordt gemaakt of daar in ieder geval geen bezwaar tegen heeft. Hiervan kan sprake zijn in persoonlijke relaties, waarbij partners van elkaar grappige deepfake filmpjes maken en die op een persoonlijke website publiceren. Het kan ook gaan om burgers die een contractuele relatie met elkaar aangaan; bijvoorbeeld een influencer die een handige kennis inhuurt om van hem deepfake filmpjes te maken.

Hierbij moeten evenwel twee zaken worden opgemerkt.

Enerzijds moet het hier gaan om expliciete, voorafgaande en bij voorkeur schriftelijke toestemming, waarbij het datasubject alle relevante informatie heeft alvorens zijn toestemming te geven (zie voor de informatieplicht paragraaf 3.2.3).[221] Het kan dus niet gaan om toestemming die impliciet wordt gegeven of verondersteld of om toestemming die achteraf wordt gegeven. Het feit dat een datasubject geen bezwaar maakt tegen een deepfake die is vervaardigd zodra hij hier weet van heeft mag ook niet worden geïnterpreteerd als stilzwijgende of impliciete toestemming.[222] Toestemming is 'elke vrije, specifieke, geïnformeerde en ondubbelzinnige wilsuiting waarmee de betrokkene door middel van een verklaring of een ondubbelzinnige actieve handeling hem betreffende verwerking van persoonsgegevens aanvaardt'.[223] De bewijslast ligt bij de verwerkingsverantwoordelijke om aan te tonen dat het datasubject zijn expliciete toestemming heeft gegeven. Als het datasubject inderdaad zijn toestemming heeft gegeven voor een deepfake, dan geldt dat hij die te allen tijde weer mag intrekken. In dat geval is de initiële, op toestemming gebaseerde, deepfake legitiem, maar moet die in principe worden verwijderd zodra de toestemming is ingetrokken, tenzij de derde en laatste voor deepfakes relevante legitieme verwerkingsgrond die hieronder aan bod zal komen soelaas kan bieden.

Anderzijds is er discussie in hoeverre ouders toestemming kunnen geven voor het verwerken

van persoonsgegevens van hun kinderen. Al langer is het sommige kinderen die adolescent of volwassen worden een doorn in het oog dat hun ouders allerhande foto's, filmpjes en andere data over hen vrijelijk beschikbaar hebben gesteld op het internet. Ouders kunnen namens een kind toestemming geven voor de verwerking van persoonsgegevens, maar dat kan slechts in zoverre dat in het belang van het kind is. Al vaker heeft de rechter kritisch geoordeeld over dergelijke praktijken en onder meer gesteld dat niet duidelijk is welk belang van een minderjarig kind wordt gediend met de publicatie van jeugdfoto's.[224] Dergelijke redenatie zal in veel gevallen ook van toepassing zijn op deepfakes gemaakt door ouders van hun eigen kinderen, met name als die in grotere groepen zijn gedeeld of via het internet vrijelijk beschikbaar zijn gemaakt.

Bij de contractuele relatie moet worden opgemerkt dat dit in puur burger-burger relaties zelden van toepassing zal zijn. Als het al een rol speelt, dan zal doorgaans een van de partijen in een professionele hoedanigheid fungeren. Denk daarbij aan een makelaar die een deepfake filmpje van een huis van een klant maakt, een keukenleverancier die middels een deepfake toont hoe de nieuwe keuken eruit komt te zien of aan een sekswerker die een deepfake van zichzelf genereert om zijn online klanten van dienst te zijn. Tussen burgers onderling zal niet snel een rechtsgeldig contract worden gesloten waar het verwerken van persoonsgegevens door middel van een deepfake noodzakelijkerwijs uit voortvloeit.

De derde en laatste relevante legitieme verwerkingsgrond voor deepfakes in horizontale verhoudingen betreft het gerechtvaardigde belang van de verwerkingsverantwoordelijke. Een verantwoordelijke zou zich daar bijvoorbeeld op kunnen beroepen als hij een geluidsfragment van iemand opneemt met als doel om daar vervolgens een deepfake van te kunnen maken. De verwerking van persoonsgegevens in zo'n geval kan legitiem zijn als 'de verwerking is noodzakelijk voor de behartiging van de gerechtvaardigde belangen van de verwerkingsverantwoordelijke of van een derde, behalve wanneer de belangen of de grondrechten en de fundamentele vrijheden van de betrokkene die tot bescherm.'[225] Een overweging bij de AVG geeft nadere invulling. 'Een dergelijk gerechtvaardigd belang kan bijvoorbeeld aanwezig zijn wanneer sprake is van een relevante en passende verhouding tussen de betrokkene en de verwerkingsverantwoordelijke, in situaties waarin de betrokkene een klant is of in dienst is van de verwerkingsverantwoordelijke. In elk geval is een zorgvuldige beoordeling geboden om te bepalen of sprake is van een gerechtvaardigd belang, alsook om te bepalen of een betrokkene op het tijdstip en in het kader van de verzameling van de persoonsgegevens redelijkerwijs mag verwachten dat verwerking met dat doel kan plaatsvinden. De belangen en de grondrechten van de betrokkene kunnen met name zwaarder wegen dan het belang van de verwerkingsverantwoordelijke wanneer persoonsgegevens worden verwerkt in omstandigheden waarin de betrokkenen redelijkerwijs geen verdere verwerking verwachten. Aangezien het aan de wetgever staat om de rechtsgrond voor persoonsgegevensverwerking door overheidsinstanties te creëren, mag die rechtsgrond niet van toepassing zijn op de verwerking door overheidsinstanties in het kader van de uitvoering van hun taken. De verwerking van persoonsgegevens die strikt noodzakelijk is voor fraudevoorkoming is ook een gerechtvaardigd

belang van de verwerkingsverantwoordelijke in kwestie. De verwerking van persoonsgegevens ten behoeve van direct marketing kan worden beschouwd als uitgevoerd met het oog op een gerechtvaardigd belang.'[226]

Er is zowel binnen de wetenschappelijke literatuur als tussen de diverse handhavende organisaties discussie over hoe streng of soepel deze bepaling moet worden geïnterpreteerd. Daarop zal hier niet uitgebreid worden ingegaan. De volgende punten zijn in ieder geval evident. Allereerst moet het gerechtvaardigd belang van de verantwoordelijke een legitiem belang zijn. Dat betekent dat het gebruik van deepfakes voor illegitieme doeleinden – stalking, afpersing, fraude, etc. – nimmer gerechtvaardigd is onder deze verwerkingsgrond (en overigens ook niet onder een van de andere gronden). Ten tweede dat de belangen van minderjarigen hier extra zwaar tellen. In principe geldt dus dat het belang van het kind om zijn gegevens niet verwerkt te zien worden bijna altijd boven het belang van de verantwoordelijke om dat wel te doen zal gaan, alhoewel ook hierop uitzonderingen voor te stellen zijn, zoals ouders die voor een kinderfeest een deepfake van hun eigen kind maken waarin het lijkt alsof dat net als superman kan vliegen. Ten derde geldt ook bij volwassen datasubjecten dat hun belangen moeten worden meegenomen. Er moet in zo'n geval worden beoordeeld welke van de belangen – die van de verantwoordelijke of die van het datasubject – gewichtiger is. Alhoewel het zeker zo zal zijn dat bij grappige deepfakes die voornamelijk ten doel hebben satire te bedrijven er geen grote belangen van het datasubject op het spel staan, is het omgekeerde ook waar: er staan geen grote belangen voor de verwerkingsverantwoordelijke op het spel om de deepfake wel te maken. Omdat hier tevens moet worden meegenomen dat de beeltenis van personen worden gemanipuleerd en dit op zichzelf al als aantasting van een mensenrecht kan worden gezien (zie paragraaf 3.3), ligt het voor de hand dat de belangen van de verwerkingsverantwoordelijke slechts in een beperkt aantal gevallen boven die van het datasubject zullen gaan.

Wel geldt dat ook satire op zich een gerechtvaardigd belang kan zijn en dat de bepaling in de Verordening niet alleen spreekt over de belangen van de verwerkingsverantwoordelijke, maar ook rept van de belangen van derden. Een derde is volgens de AVG 'een natuurlijke persoon of rechtspersoon, een overheidsinstantie, een dienst of een ander orgaan, niet zijnde de betrokkene, noch de verwerkingsverantwoordelijke, noch de verwerker, noch de personen die onder rechtstreeks gezag van de verwerkingsverantwoordelijke of de verwerker gemachtigd zijn om de persoonsgegevens te verwerken'.[227] Dit betekent dat derden kunnen zijn personen die een deepfake bekijken. Hun belang, bijvoorbeeld in het geval van een satirische deepfake, zou dan ook moeten worden meegewogen. Dat zou kunnen betekenen dat als er een miljoen personen zeer veel plezier beleven aan een hilarische deepfake, hun belangen uiteindelijk boven de belangen van het datasubject in kwestie zullen gaan. Dit zal vermoedelijk evenwel niet opgaan bij seksuele deepfakes. Het utilistische argument dat alhoewel een vrouw (het datasubject) last heeft van een pornografische deepfake, maar haar belangen niet opwegen tegen het plezier dat duizenden mannen er aan beleven zal naar verwachting niet worden gehonoreerd door een rechter, te meer daar het feit dat meer mensen toegang hebben tot een schadelijke deepfake ook kan bijdragen aan het belang van het datasubject om bezwaar te maken tegen een verwerking.

Bovenstaande beschrijving ging uit van het verwerken van persoonsgegevens op basis van het doel om een deepfake te maken. Daarvoor kan toestemming worden verkregen of dat kan in voorkomende gevallen gelegitimeerd zijn als het belang van de verwerkingsverantwoordelijke en eventuele derden boven die van het datasubject gaan. Een ander geval is, zoals in paragraaf 3.2.4 beschreven, wanneer persoonsgegevens voor het ene doel zijn verzameld en ze vervolgens voor een ander doel, namelijk het maken van een deepfake, worden gebruikt. De AVG stelt dat het gebruiken van gegevens voor nieuwe doeleinden mag als daarvoor toestemming van het datasubject is verkregen of als dit een publiek belang dient. 'Wanneer de betrokkene zijn toestemming heeft gegeven of wanneer de verwerking gebaseerd is op Unierecht of lidstatelijk recht dat in een democratische samenleving een noodzakelijke en evenredige maatregel vormt voor met name het waarborgen van belangrijke doelstellingen van algemeen belang, moet de verwerkingsverantwoordelijke de mogelijkheid hebben de persoonsgegevens verder te verwerken, ongeacht of dat verenigbaar is met de doeleinden. In ieder geval dient ervoor te worden gezorgd dat de in deze verordening vervatte beginselen worden toegepast en dat de betrokkene met name wordt geïnformeerd over dergelijke andere doeleinden en over zijn rechten, waaronder het recht om bezwaar te maken.'[228]

Als A derhalve aan B een foto heeft gestuurd met als doel een vakantie-fotoboek te maken en B vraag een jaar later aan A of die er akkoord mee is dat de foto ook voor een deepfake wordt gebruikt en A geeft daarop expliciet zijn toestemming, dan is die tweede toestemming de verwerkingsgrond voor het vervaardigen van de deepfake en niet de eerste. Zoals uit de laatste zin van het hierboven geciteerde blijkt zal bij een verwerking van persoonsgegevens voor een nieuw doel hoe dan ook worden voldaan aan de transparantieplicht als beschreven in paragraaf 3.2.4. In feite gaat de AVG dus ervanuit dat er een nieuwe gegevensverwerking plaatsvindt, waardoor de hele Verordening op dat moment weer langs moeten worden gelopen en alle rechten en plichten van toepassing zijn op die verwerking als zijnde een nieuwe oorspronkelijke verwerking.

Dan is nog de vraag of het gebruiken van data voor een ander doel ook mag als een nieuwe verwerkingsgrondslag wordt gevonden in het 'gerechtvaardigd belang' van de verwerkingsverantwoordelijke of de belangen van derden. Dit is met name relevant in het geval de verantwoordelijke en het datasubject niet in directe relatie tot elkaar staan, zoals wanneer er deepfakes van Bekende Nederlanders worden gemaakt en ten behoeve daarvan materiaal van het internet wordt gehaald. Voor dat geval stelt de AVG het volgende: 'Wanneer de verwerking voor een ander doel dan dat waarvoor de persoonsgegevens zijn verzameld niet berust op toestemming van de betrokkene of op een Unierechtelijke bepaling of een lidstaatrechtelijke bepaling die in een democratische samenleving een noodzakelijke en evenredige maatregel vormt ter waarborging van de in artikel 23, lid 1, bedoelde doelstellingen houdt de verwerkingsverantwoordelijke bij de beoordeling van de vraag of de verwerking voor een ander doel verenigbaar is met het doel waarvoor de persoonsgegevens aanvankelijk zijn verzameld onder meer rekening met: a) ieder verband tussen de doeleinden waarvoor de persoonsgegevens zijn verzameld, en de doeleinden van de voorgenomen verdere verwerking; b) het kader

waarin de persoonsgegevens zijn verzameld, met name wat de verhouding tussen de betrokkenen en de verwerkingsverantwoordelijke betreft; c) de aard van de persoonsgegevens, met name of bijzondere categorieën van persoonsgegevens worden verwerkt, overeenkomstig artikel 9, en of persoonsgegevens over strafrechtelijke veroordelingen en strafbare feiten worden verwerkt, overeenkomstig artikel 10; d) de mogelijke gevolgen van de voorgenomen verdere verwerking voor de betrokkenen; e) het bestaan van passende waarborgen, waaronder eventueel versleuteling of pseudonimisering.'[229]

Het lijkt er daarom op dat hergebruik in principe niet toegestaan, tenzij (1) er een nieuwe verwerkingsgrondslag wordt gevonden in de zin van toestemming of een publiek belang of (2) het nieuwe gebruik van de data een doel dient dat verenigbaar is met het oorspronkelijke doel. Dat lijkt ook te volgen uit een overweging uit de AVG. 'Om na te gaan of een doel van verdere verwerking verenigbaar is met het doel waarvoor de persoonsgegevens aanvankelijk zijn verzameld, moet de verwerkingsverantwoordelijke, nadat hij aan alle voorschriften inzake rechtmatigheid van de oorspronkelijke verwerking heeft voldaan, onder meer rekening houden met: een eventuele koppeling tussen die doeleinden en de doeleinden van de voorgenomen verdere verwerking; het kader waarin de gegevens zijn verzameld; met name de redelijke verwachtingen van de betrokkenen op basis van hun verhouding met de verwerkingsverantwoordelijke betreffende het verdere gebruik ervan; de aard van de persoonsgegevens; de gevolgen van de voorgenomen verdere verwerking voor de betrokkenen; en passende waarborgen bij zowel de oorspronkelijke als de voorgenomen verdere verwerkingen.'[230]

Of aan deze voorwaarden is voldaan zal van geval tot geval moeten worden beoordeeld.

3.2.8 Legitieme verwerking bijzondere persoonsgegevens

Voor de verwerking van zogenoemde 'bijzondere persoonsgegevens' geldt binnen de AVG een nee-tenzij regime. Bijzondere persoonsgegevens zijn gegevens waaruit ras of etnische afkomst, politieke opvattingen, religieuze of levensbeschouwelijke overtuigingen, of het lidmaatschap van een vakbond blijken; ook gaat het om de verwerking van genetische gegevens, biometrische gegevens met het oog op de unieke identificatie van een persoon, of gegevens over gezondheid, of gegevens met betrekking tot iemands seksueel gedrag of seksuele gerichtheid en verwerking van strafrechtelijke gegevens. Daarbij is van belang dat een vrij brede benadering wordt gehanteerd ten aanzien van wat een 'bijzonder persoonsgegeven' is en wat niet. Zo gaat het bij gezondheidsgegevens niets slechts om ernstige ziekten, maar bleek uit de eerdergenoemde zaak van Bodil Lindqvist dat ook informatie over iemand met een gebroken been hieronder valt. Dit betekent dat niet alleen voor zover er seksuele zaken uit deepfakes blijken of strafrechtelijke handelingen lijken te worden verricht, dit bijzondere regime uit de AVG van toepassing zal zijn, maar dat dit veel vaker het geval zal zijn.

De AVG geeft een aantal uitzonderingsgronden voor het verbod op de verwerking van bijzondere persoonsgegevens. Daarvan zijn er twee van belang voor deepfakes in horizontale relaties.

Allereerst betreft het de toestemming van het datasubject. Daarbij gelden alle eerdere besproken voorwaarden, plus nog een

additionele verzwaring nu het in dit geval moet gaan om *uitdrukkelijke* toestemming. Interessant is dat de AVG op dit punt bepaalt dat lidstaten, in dit geval Nederland, ervoor kan kiezen om toestemming voor bepaalde verwerkingen van bijzondere verwerkingen niet is toegestaan. Daar heeft de Nederlandse wetgever echter geen gebruik van gemaakt. Wellicht kan hier in de toekomst aanleiding toe zijn, bijvoorbeeld als het gaat om werkgevers die werknemers om toestemming vragen om een deepfake van hen te maken, waarbij het zeker niet in de laatste plaats zal gaan om afbeeldingen van sekswerkers. Hoe dan ook blijkt uit de toevoeging van de *uitdrukkelijk* gegeven toestemming dat hier nog preciezer zal worden gekeken naar de door het datasubject gegeven toestemming en dat deze toestemming noch evidenter moet zijn dan al het geval is bij toestemming als grond voor het verwerken van gewone persoonsgegevens.

Ten tweede kan het gaan om een het geval 'de verwerking heeft betrekking op persoonsgegevens die kennelijk door de betrokkene openbaar zijn gemaakt'. De Nederlandse wetgever heeft deze grond ook overgenomen in de UAVG en verklaarde in de Memorie van Toelichting omtrent deze verwerkingsgrond: 'Evenals bij onderdeel a [toestemming van het datasubject] ligt de rechtvaardigingsgrond voor de uitzondering besloten in het handelen of het gedrag van de betrokkene zélf. Anders dan bij onderdeel a is er echter geen sprake van op de gegevensverwerking gerichte toestemming, maar van een spontane gedraging van de betrokkene waar niet door enig ander persoon met het oog op een eventuele gegevensverwerking om is gevraagd. Dat de gegevens openbaar zijn, moet derhalve volgen uit gedrag van de betrokkene waaruit de intentie om openbaar te maken uitdrukkelijk blijkt. Van vrijwillige openbaarmaking kan ook impliciet sprake zijn. Een voorbeeld van een dergelijke vrijwillige openbaarmaking zijn gegevens in een telefoongids, indien daarin ook gevoelige gegevens zouden voorkomen. Ieder is immers vrij de vermelding daarin te voorkomen. Indien een bepaald persoonsgegeven openbaar is, maar niet uit vrije wil van de betrokkene, dan mag het gegeven niet op grond onderdeel d worden verwerkt.'[231]

Hieruit lijken twee belangrijke zaken te volgen in het kader van deepfakes.

Enerzijds gaat deze uitzonderingsgrond reeds uit van hergebruik. Er staan bijvoorbeeld gegevens op het internet die kennelijk openbaar zijn gemaakt door het datasubject, bijvoorbeeld selfies op zijn Insta-pagina, en die worden vervolgens verder verwerkt door een andere burger. De vraag is hierbij of het alsnog in principe moet gaan om een doeleinde voor het hergebruik die niet onverenigbaar is met het doel van de oorspronkelijke publicatie. Als dat het geval is, dan zal slechts in het geval de gegevens openbaar zijn gemaakt ten behoeve van het maken van deepfakes of breder, doeleinden die niet met het doeleinde voor het maken van een deepfake onverenigbaar zijn, toegestaan zijn om deze gegevens te hergebruiken. Als dat niet het geval is, dan heeft dat het merkwaardige gevolg dat er minder strenge eisen lijken te gelden voor het hergebruik van bijzondere persoonsgegevens dan voor het hergebruik van gewone persoonsgegevens. Noch de AVG noch de parlementaire wetsgeschiedenis geven uitsluitsel over de juiste interpretatie.

Anderzijds benadrukt de regering dat het ook kan gaan om 'impliciete' goedkeuring, wat onder

meer kan worden afgeleid uit het feit dat een datasubject geen gebruik maakt van het recht om gegevensverwerking tegen te gaan. Hier lijkt een redenatie te worden gevolgd die ten gevolg kan hebben dat er minder strenge regels gelden voor het verwerken van bijzondere persoonsgegevens dan voor gewone persoonsgegevens, aangezien dergelijke impliciete toestemming juist expliciet wordt afgewezen als toestemming wordt gebruikt als legitieme verwerkingsgrond voor gewone persoonsgegevens. Ook is hierbij de vraag hoe ver de redenatie van de regering strekt. Datasubjecten hebben immers altijd het recht om te verzoeken gegevens te verwijderen of te stoppen met verdere verwerking (alhoewel dat verzoek niet altijd hoeft te worden ingewilligd, zie de hiernavolgende paragraaf). Moet, de lijn van de regering volgend, worden aangenomen dat data waarvan het datasubject weet dat zij openbaar zijn gemaakt maar ten aanzien waarvan hij geen gebruik maakt van zijn recht op bezwaar of beperking, kunnen worden verwerkt op basis van impliciete goedkeuring? Waarschijnlijk moet een beperktere uitleg aan de bepaling worden gegeven, maar eenduidigheid is er in de literatuur of jurisprudentie niet op dit punt. [232]

3.2.9 Recht om vergeten te worden

Dan geldt er nog een aantal rechten van het datasubject. Twee daarvan zijn al aan de orde gekomen, namelijk het recht om informatie te vragen aan de verantwoordelijke aangaande de gegevensverwerking, waaronder ook valt het recht op inzage van de gegevens en het recht op een kopie van die gegevens, en het recht op rectificatie en aanvulling van incorrecte en/of onvolledige gegevens. Daarnaast geldt dat het datasubject mag verzoeken tijdelijk met de verwerking te stoppen, bijvoorbeeld totdat de incorrecte gegevens op zijn verzoek zijn gecorrigeerd.[233] Daarnaast is het recht op bezwaar relevant indien een verwerkingsverantwoordelijke een beroep zou doen op zijn 'gerechtvaardigde belang' als legitieme verwerkingsgrond (paragraaf 3.2.7). Het datasubject mag dan altijd betwisten dat het om een legitiem belang van de verwerkingsverantwoordelijke gaat en/of dat belang hoger staat dan zijn eigen belang om zijn persoonsgegevens niet te verwerken. 'De verwerkingsverantwoordelijke staakt de verwerking van de persoonsgegevens tenzij hij dwingende gerechtvaardigde gronden voor de verwerking aanvoert die zwaarder wegen dan de belangen, rechten en vrijheden van de betrokkene of die verband houden met de instelling, uitoefening of onderbouwing van een rechtsvordering.'[234] Dit betekent dat de verwerkingsverantwoordelijke mag weigeren om aan het verzoek gehoor te geven als hij meent dat zijn belangen wel degelijk voor die van het datasubject moeten gaan. In het geval beide partijen ook na verder overleg hun oorspronkelijke mening blijven toegedaan zal de zaak uiteindelijk door het datasubject aan de AP of een rechter moeten worden voorgelegd.

Belangrijk is ook nog kort te verwijzen naar wat in de volksmond het recht op vergetelheid is komen te heten. Hierover is veel commotie ontstaan, maar in wezen in het een recht met zeer beperkte reikwijdte en van zeer beperkt belang, ook in het kader van deepfakes. Het recht stelt simpelweg dat als de persoonsgegevens van een datasubject in strijd met de AVG worden verwerkt, hij mag verzoeken met de verwerking te stoppen. Zelfs hierop zijn echter nog uitzonderingen, namelijk in het geval de verwerking noodzakelijk is in het kader van de vrijheid van meningsuiting. Daarop zal een verwerkingsverantwoordelijk vaak een beroep kunnen doen, aangezien die

vrijheid een zeer brede reikwijdte heeft en onder meer omvat satire bedrijven en het publiek of de samenleving te schokken (zie uitgebreider paragraaf 3.3). Waarschijnlijk vallen hier echter niet onder duidelijk illegale activiteiten, zoals het gebruiken van deepfakes voor fraude, wraakporno of oplichting.[235]

3.2.10 Vrijheid van meningsuiting

Tot slot is voor privacyschendingen in horizontale verhoudingen nog relevant dat het recht op gegevensbescherming kan botsen met andere grondrechten (dit geldt ook ten aanzien van het recht op privacy, als dat wordt toegepast binnen horizontale verhoudingen). Bij bedrijven valt daarbij te denken aan het recht op ondernemerschap (waaronder eenmanszaken), zoals onder meer is vervat in artikel 16 van het Handvest van de Fundamentele Rechten van de Europese Unie: "De vrijheid van ondernemerschap wordt erkend overeenkomstig het recht van de Unie en de nationale wetgevingen en praktijken."[236]

In burger-burger relaties zal het daarbij met name gaan om het recht op vrijheid van meningsuiting, zoals onder meer neergelegd in artikel 11 van dat Handvest: "1. Eenieder heeft recht op vrijheid van meningsuiting. Dit recht omvat de vrijheid een mening te hebben en de vrijheid kennis te nemen en te geven van informatie of ideeën, zonder inmenging van enig openbaar gezag en ongeacht grenzen. 2. De vrijheid en de pluriformiteit van de media worden geëerbiedigd." Onder het recht op de vrijheid van meningsuiting valt ook het recht op het vergaren en het verspreiden van informatie.

De AVG geeft aan dat landen nadere regels kunnen stellen ten aanzien van de verwerking van persoonsgegevens voor dergelijke doeleinden. Opvallend is dat Nederland er in de UAVG voor heeft gekozen om slechts voor journalistieke werkzaamheden uitzonderingen op het gegevensbeschermingsrecht neer te leggen en niet voor activiteiten in het kader van vrijheid van meningsuiting in het algemeen.

AVG

Artikel 85 Verwerking en vrijheid van meningsuiting en van informatie

1. De lidstaten brengen het recht op bescherming van persoonsgegevens overeenkomstig deze verordening wettelijk in overeenstemming met het recht op vrijheid van meningsuiting en van informatie, daaronder begrepen de verwerking voor journalistieke doeleinden en ten behoeve van academische, artistieke of literaire uitdrukkingsvormen.
2. Voor verwerking voor journalistieke doeleinden of ten behoeve van academische, artistieke of literaire uitdrukkingsvormen stellen de lidstaten uitzonderingen of afwijkingen vast van hoofdstuk II (beginselen), hoofdstuk III (rechten van de betrokkene), hoofdstuk IV (de verwerkingsverantwoordelijke en de verwerker), hoofdstuk V (doorgifte van persoonsgegevens naar derde landen of internationale organisaties), hoofdstuk VI (onafhankelijke toezichthoudende autoriteiten), hoofdstuk VII (samenwerking en coherentie) en hoofdstuk IX (specifieke gegevensverwerkingssituaties) indien deze noodzakelijk zijn om het recht op bescherming van persoonsgegevens in overeenstemming te brengen met de vrijheid van meningsuiting en van informatie.
3. Elke lidstaat deelt de Commissie de overeenkomstig lid 2 vastgestelde wetgevingsbepalingen mee, alsook onverwijld alle latere wijzigingen daarvan.

UAVG

Artikel 43. Uitzonderingen inzake journalistieke doeleinden of academische, artistieke of literaire uitdrukkingsvormen

1. Deze wet, met uitzondering van de artikelen 1 tot en met 4 en 5, eerste en tweede lid, is niet van toepassing op de verwerking van persoonsgegevens voor uitsluitend journalistieke doeleinden en ten behoeve van uitsluitend academische, artistieke of literaire uitdrukkingsvormen.
2. De navolgende hoofdstukken en artikelen van de verordening zijn niet van toepassing op de verwerking van persoonsgegevens voor uitsluitend journalistieke doeleinden en ten behoeve van academische, artistieke of literaire uitdrukkingsvormen: a.artikel 7, derde lid, en artikel 11, tweede lid: b.hoofdstuk III; c.hoofdstuk IV, met uitzondering van de artikelen 24, 25, 28, 29 en 32; d.hoofdstuk V; e.hoofdstuk VI; en f.hoofdstuk VII.
3. De artikelen 9 en 10 van de verordening zijn niet van toepassing voor zover de verwerking van de in die artikelen bedoelde gegevens noodzakelijk is voor het journalistieke doel of de academische, artistieke of literaire uitdrukkingsvorm.

Figuur 18: Vergelijking AVG & UAVG op VVMU

In hoeverre er dus sprake kan zijn van een uitzondering voor het verwerken van persoonsgegevens door burgers voor niet journalistieke werkzaamheden die desalniet–temin zijn te kwalificeren als de verwerking van persoonsgegevens in het kader van de vrijheid van meningsuiting, blijft onduidelijk. Wel is duidelijk dat burgers zich in een conflict waarbij de ene burger ongewenst persoonsgegevens verzamelt van de andere burger, zich zullen beroepen op het grondwettelijke recht op gegevensbescherming enerzijds en het grondwettelijke recht op vrijheid van meningsuiting anderzijds. Daarbij zullen zij verwijzen naar ofwel de Nederlandse Grondwet, ofwel het Handvest van de grondrechten van de Europese Unie, ofwel het Europees Verdrag voor de Rechten van de Mens. Via die lijn zou dan eventueel alsnog een beperking ten aanzien van de gegevensbeschermingsregels kunnen worden afgedwongen (dat wil zeggen, buiten de door de UAVG voorziene omstandigheden). Voor journalisten geldt dat zij, voorover zij persoonsgegevens verwerken voor journalistieke doeleinden, ontheven kunnen zijn van een groot aantal van de verplichtingen die de AVG aan hen oplegt.

Toch benadrukt de Leidraad van de Raad voor de Journalistiek expliciet dat journalisten te alle tijden zijn gehouden aan het respect voor privacy: “In een publicatie mag de privacy van personen niet verder worden aangetast dan in het kader van de berichtgeving redelijkerwijs noodzakelijk is. Een inbreuk op de privacy is onzorgvuldig wanneer deze niet in redelijke verhouding staat tot het maatschappelijk belang van de publicatie. Journalisten publiceren geen foto’s en zenden geen beelden uit die zijn gemaakt van personen in niet algemeen toegankelijke ruimten zonder hun toestemming, en gebruiken evenmin brieven en persoonlijke aantekeningen zonder toestemming van betrokkenen. Journalisten mogen personen niet langdurig lastigvallen, hinderlijk volgen of schaduwen. Journalisten dienen te voorkomen dat informatie of beelden worden gepubliceerd waardoor verdachten en veroordeelden door het grote publiek eenvoudig kunnen worden geïdentificeerd en getraceerd.”[237] Ook journalisten zijn dus, voor zover dat hun werkzaamheden niet ondermijnt, gehouden aan privacy- en gegevensbeschermingsprincipes. Dat geldt in principe ook voor burgerjournalisten.

3.2.11 Conclusie

Deepfakes lopen tegen een aantal obstakels op 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, wat ook 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. Wel moet worden bedacht dat er uitzonderingen kunnen bestaan in de vorm van de huishoudelijke exceptie en de verwerking van gegevens in het kader van de vrijheid van meningsuiting. Hoe nauw of wijd deze uitzonderingen dienen te worden geïnterpreteerd in de context van deepfakes is echter niet op voorhand en in het algemeen vast te stellen.

3.3 Vrijheid van meningsuiting en de bescherming van eer en goede naam

De vorige paragraaf sloot af met een verwijzing naar de vrijheid van meningsuiting, het recht dat kan worden ingeroepen door degene die de deepfake vervaardigd. Daarom is het van belang niet alleen kort stil te staan bij dit recht, zoals volgt uit artikel 10 van het Europees Verdrag voor de Rechten van de Mens, een instrument van de Raad van Europa waarop het Europees Hof voor de Rechten van de Mens toeziet, maar ook met het recht op privacy, artikel 8 van datzelfde Verdrag, dat een bredere reikwijdte heeft dan het recht op gegevensbescherming, zoals dat is neergelegd in de Algemene Verordening Gegevensbescherming van de EU. In deze paragraaf zal kort worden stilgestaan bij de botsing tussen deze twee rechten en hoe de jurisprudentie van het EHRM op deepfakes van toepassing zouden kunnen zijn. Dat zal worden gedaan aan de hand van vier stappen. De eerste twee stappen betreffen de reikwijdte van het recht op privacy. Ten eerste zal worden stilgestaan bij het feit dat Artikel 8 EVRM ook ziet op het verzamelen van persoonsgegevens, juist als dit in de privésfeer wordt gedaan, maar zelfs in werkrelaties of in de publieke sfeer. Ten tweede zal worden aangetoond dat het EHRM onder dit recht steeds meer een recht op de bescherming van eer en goede naam is gaan scharen. Ten derde zal kort worden aangegeven wat de reikwijdte van het recht op vrijheid van meningsuiting is. Tot vierde zal met name worden besproken hoe het EHRM omgaat met een botsing van deze twee rechten. Daarbij zal met name worden ingegaan op de bijzondere positie van publieke figuren, waarvan immers ook veel deepfakes worden gemaakt. Tot slot wordt een korte conclusie gegeven.

3.3.1 Persoonsgegevens en de reasonable expectation of privacy

Artikel 8 Recht op eerbiediging van privé-, familie- en gezinsleven

1. Een ieder heeft recht op respect voor zijn privé leven, zijn familie- en gezinsleven, zijn woning en zijn correspondentie.

2. Geen inmenging van enig openbaar gezag is toegestaan in de uitoefening van dit recht, dan voor zover bij de wet is voorzien en in een democratische samenleving noodzakelijk is in het belang van de nationale veiligheid, de openbare veiligheid of het economisch welzijn van het land, het voorkomen van wanordelijkheden en strafbare feiten, de bescherming van de gezondheid of de goede zeden of voor de bescherming van de rechten en vrijheden van anderen.

Het recht op privacy en het recht op gegevensbescherming hebben een verschillende achtergrond, traditie en grondgedachte. Het recht op gegevensbescherming is van recentere oorsprong en valt in veel opzichten samen met de opkomst van moderne technologieën waarmee het verzamelen, opslaan en verwerken van persoonsgegevens steeds gemakkelijker werd. Persoonsgegevens betreffen niet alleen privé- of privacygevoelige gegevens. Het kan ook gaan om openbare en niet-gevoelige informatie, als deze zouden kunnen worden gebruikt om iemand te identificeren. Even ancillary information, such as "the man wearing a black suit" may identify someone out of the passers-by standing at a traffic light.'[238] Het recht op privacy gaat traditioneel slechts om de bescherming van gegevens voor zover die het privéleven van een persoon raken of worden beschermd door de vertrouwelijkheid van communicatie. Alhoewel het EHRM in de loop der jaren bereid is geweest steeds meer type data en zelfs metadata onder de materiele reikwijdte van artikel 8 EVRM te brengen, is het recht op privacy minder breed dan het recht op gegevensbescherming.[239] De beelden en/of geluiden die in deepfakes te zien of te horen zijn zullen in de meeste gevallen ook onder de reikwijdte van het recht op privacy vallen, omdat het EHRM in het verleden slechts anders heeft geoordeeld over zeer ongevoelige gegevens of alledaagse gegevensverwerkingen.[240]

Daarbij is in het kader van deepfakes, zeker gezien de huishoudelijke exceptie, het volgende van belang. Het recht op privacy is juist van toepassing als gegevens in de privésfeer worden verzameld en in mindere mate als het gaat om de gegevensverzameling in de publieke ruimte. In zekere zin lijkt het dus het spiegelbeeld van het gegevensbeschermingsrecht. Toch geldt ook in de publieke ruimte en in werkcontext het recht op privacy, zelfs als mensen weten dat hun gegevens worden verzameld of zij die zelf openbaar hebben gemaakt. Vaak wordt ook in Europees verband verwezen naar de Amerikaanse 'reasonable expectation of privacy' doctrine, waaruit volgt dat in de publieke ruimte geen of weinig privacy te verwachten valt, maar dat is slechts ten dele op zijn plaats. Alhoewel het EHRM de term in iets meer dan een dozijn (ten opzichte van vele duizenden uitspraken waarin het die term niet gebruikt) rechtszaken toepast, kiest het voor een geheel eigen doctrine die hier de moeite van het bespreken waard is, niet alleen vanwege het feit dat artikel 8 EVRM zowel in de privé als in de publieke sfeer geldt, maar ook om twee additionele punten. Ten eerste dat ook als personen gevoelige data expliciet openbaar hebben gemaakt zij een recht op privacy hebben. Ten tweede geldt dat zelfs voor onrechtmatig gedrag.

De eerste keer dat de doctrine van de "redelijke verwachting" werd gebruikt in een zaak voor het EHRM was door de Britse regering in de ontvankelijkheidsbeslissing Halford v. UK (1995), waarin de verzoekster beweerde dat zij als gevolg van haar klachten over discriminatie op grond van geslacht was onderworpen aan toezicht, waaronder het afluisteren van haar kantoor en het onderscheppen van haar gesprekken op haar privételefoon thuis en haar kantoortelefoons. Verzoekster had twee telefoons in haar kantoor: een telefoon met een extern nummer voor persoonlijke gesprekken en een telefoon voor politiewerk. De gesprekken met deze beide telefoons werden betaald door de politie, haar werkgever. Bijgevolg betoogde de regering dat 'the applicant had no reasonable expectation of privacy in relation to those telephones.'[241] Het Hof

(1997) benadrukte echter dat het bedrijfspand onder omstandigheden iemands 'woning' kan vormen, zoals beschermd op grond van artikel 8, lid 1, EVRM, dat 'privéleven', een andere term die in dat lid wordt genoemd, zich ook kan afspelen op iemands werkplek, en dat 'correspondentie' zowel communicatie vanaf privé- als vanaf zakelijke telefoons omvat.[242]

'There is no evidence of any warning having been given to Ms Halford, as a user of the internal telecommunications system operated at the Merseyside police headquarters, that calls made on that system would be liable to interception. She would, the Court considers, have had a reasonable expectation of privacy for such calls, which expectation was moreover reinforced by a number of factors. As Assistant Chief Constable she had sole use of her office where there were two telephones, one of which was specifically designated for her private use. Furthermore, she had been given the assurance, in response to a memorandum, that she could use her office telephones for the purposes of her sex-discrimination case. For all of the above reasons, the Court concludes that the conversations held by Ms Halford on her office telephones fell within the scope of the notions of "private life" and "correspondence" and that Article 8 is therefore applicable to this part of the complaint.'[243]

In de zaak P.G. en J.H. tegen het Verenigd Koninkrijk (2001) ging het Hof nog een stap verder toen verzoekers klaagden dat er afluisterapparatuur was gebruikt terwijl zij op het politiebureau waren om audio-opnames te maken. Het Hof beklemtoonde nadrukkelijk dat 'a person's reasonable expectations as to privacy may be a significant, although not necessarily conclusive' factor en benadrukte dat mensen ook in openbare ruimten privacy hebben, vooral wanneer er systematische of permanente registraties van hen worden gemaakt.[244] Vervolgens werd in de zaak Copland tegen het Verenigd Koninkrijk (2007) het telefoon-, e-mail- en internetgebruik van verzoekster gecontroleerd door de werkgever. Het Hof beklemtoonde dat verzoekster 'had been given no warning that her calls would be liable to monitoring, therefore she had a reasonable expectation as to the privacy of calls made from her work telephone. The same expectation should apply in relation to the applicant's e-mail and Internet usage.'[245]

In de zaak Peev tegen Bulgarije (2007) klaagde de verzoeker dat zijn kantoor was doorzocht en dat zijn ontwerp-ontslagbrief in beslag was genomen. Het Hof was van oordeel dat de verzoeker een redelijke privacyverwachting had, 'if not in respect of the entirety of his office, at least in respect of his desk and his filing cabinets. This is shown by the great number of personal belongings that he kept there. Moreover, such an arrangement is implicit in habitual employer-employee relations and there is nothing in the particular circumstances of the case – such as a regulation or stated policy of the applicant's employer discouraging employees from storing personal papers and effects in their desks or filing cabinets – to suggest that the applicant's expectation was unwarranted or unreasonable. The fact that he was employed by a public authority and that his office was located on government premises does not of itself alter this conclusion, especially considering that the applicant was not a prosecutor, but a criminology expert employed by the Prosecutor's Office. Therefore, a search which extended to the applicant's desk and filing cabinets must be regarded as

an interference with his private life.'[246] In Steeg tegen Duitsland (2008) heeft het Hof de redelijke privacyverwachting uitgebreid tot de inhoud van documenten en elektronische opslagmedia.[247] In Antovic en Mirkovic v. Montenegro (2017) breidde het EHRM zijn benadering uit tot opnames die universiteitsdocenten in klaslokalen maken, mogelijk om hun onderwijskwaliteit te beoordelen, die weliswaar niet privé, maar openbaar zijn, maar wel werkplekken zijn waar sociale relaties worden ontwikkeld.[248]

Een opmerkelijke stap werd gezet in Pay v. UK (2008), waarin de verzoeker in dienst was getreden van de Lancashire Probation Service en betrokken was bij de behandeling van zedendelinquenten. Hij was ook directeur van Roissy, een organisatie die op internet adverteerde als bouwer en leverancier van BDSM-producten en als organisator van BDSM-evenementen en -voorstellingen. Er circuleerde ook een foto van verzoeker, die een masker droeg, met twee halfnaakte vrouwen. Roissy was geregistreerd op het adres van Pay en zijn website bevatte links naar een aantal BDSM-websites waarop reclame werd gemaakt voor verschillende evenementen en waarop foto's stonden van verzoeker en anderen, halfnaakt, terwijl zij handelingen verrichtten die volgens de begeleidende tekst hadden plaatsgevonden in een plaatselijke privéclub voor leden en waarbij sprake was van mannelijke dominantie over onderdanige vrouwen. Hij werd van zijn werk ontslagen omdat zijn gedrag onverenigbaar werd geacht te zijn met zijn behandeling van zedendelinquenten. Het Hof erkende dat de aard van de handelingen blijkt uit internetfoto's en uit advertenties. Het erkent dat zijn gedrag en openheid daarover 'could give rise to doubts as to whether the applicant's activities may be said to fall with the scope of private life and, if so, whether [] there has been a waiver or forfeiture of the rights guaranteed by Article 8. The Court notes, however, that the applicant's performances took place in a nightclub which was likely to be frequented only by a self-selecting group of like-minded people and that the photographs of his act which were published on the internet were anonymised'.[249] Daarom was het EHRM bereid de zaak te beoordelen op basis van de veronderstelling dat de verzoeker zelfs in dit geval een redelijke verwachting van privacy had.

Een laatste belangrijke stap werd gezet in Benedik tegen Slovenië (2018), waarin het Hof in een controversiële zaak zijn benadering uitbreidde tot buiten de werkplek. De internetverbinding van verzoeker werd gemonitord, omdat hij kinderporno verspreidde. Opmerkelijk was dat het Hof benadrukte dat de verzoeker verwachtte dat zijn activiteiten privé zouden blijven en dat zijn identiteit niet bekend zou worden gemaakt. Hoewel het EHRM accepteerde dat de verzoeker zijn dynamische IP-adres niet verborgen hield, onderstreepte het ook dat dit niet doorslaggevend kan zijn bij de beoordeling of zijn verwachting van privacy vanuit objectief oogpunt redelijk was. Op dat punt herhaalde het Hof dat anonimiteit op het internet een belangrijk onderdeel is van het recht op privacy, dat een dynamisch IP-adres, ook al is het zichtbaar voor andere gebruikers van het netwerk, niet kan worden herleid tot de specifieke computer zonder verificatie door de ISP, om vervolgens te concluderen dat het 'sufficient to note that Article 37 of the Constitution guaranteed the privacy of correspondence and of communications and required that any interference with this right be based on a court order. Therefore, also from the standpoint of the legislation in force at the

relevant time, the applicant's expectation of privacy with respect to his online activity could not be said to be unwarranted or unreasonable.'[250]

Bovenstaande zegt niets over de uiteindelijke legitimiteit van de gegevensverwerking, maar wel over de vraag of het verzamelen van de gegevens onder de reikwijdte van het recht op privacy valt. Daarbij is van belang dat zowel het verzamelen in de privésfeer als in hoge mate het verzamelen in de werksfeer en de publieke sfeer onder dit recht zal vallen. Dat recht geldt zelfs als mensen zelf hun gegevens openbaar hebben gemaakt, zoals Pay die allerhande beeldmateriaal van zichzelf op het internet zette. Dat betekent dat als van die beelden vervolgens een deepfake zou worden gemaakt, het argument niet kan zijn dat hij niet langer een beroep op privacy kan doen. Ook blijkt dat als iemand bijvoorbeeld deepfake-kinderporno zou maken en verspreiden, hij alsnog een beroep op zijn recht op privacy kan doen.

3.3.2 Eer en goede naam

Het Europees Verdrag is gebaseerd op de Universele Verklaring voor de Rechten van de Mens (UVRM) en artikel 8 EVRM is gebaseerd op artikel 12 van het UVRM, dat luidt: 'No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honor and reputation.' Bij het opstellen van de Universele Verklaring werd de opneming ten aanzien van de bescherming van eer en goede naam reeds uitvoerig besproken, aangezien het hier voornamelijk gaat om horizontale relaties en aanvallen door of via de media. Hoewel de bescherming van deze belangen uiteindelijk in het kader van de UVRM zijn opgenomen, hebben de opstellers van het EVRM, dat zich alleen op verticale verhoudingen (tussen staat en burger) richt, de bescherming van iemands reputatie niet opgenomen als subjectief recht in het kader van het recht op eerbiediging van de persoonlijke levenssfeer. In plaats daarvan is dit aspect ondergebracht onder lid 2 van artikel 10 EVRM, dat het recht op vrijheid van meningsuiting bevat. Artikel 10 EVRM bepaalt:

> *Artikel 10 Vrijheid van meningsuiting*
>
> *1. Een ieder heeft recht op vrijheid van meningsuiting. Dit recht omvat de vrijheid een mening te koesteren en de vrijheid om inlichtingen of denkbeelden te ontvangen of te verstrekken, zonder inmenging van enig openbaar gezag en ongeacht grenzen. Dit artikel belet Staten niet radio- omroep-, bioscoop of televisieondernemingen te onderwerpen aan een systeem van vergunningen.*
>
> *2. Daar de uitoefening van deze vrijheden plichten en verantwoordelijkheden met zich brengt, kan zij worden onderworpen aan bepaalde formaliteiten, voorwaarden, beperkingen of sancties, die bij de wet zijn voorzien en die in een democratische samenleving noodzakelijk zijn in het belang van de nationale veiligheid, territoriale integriteit of openbare veiligheid, het voorkomen van wanordelijkheden en strafbare feiten, de bescherming van de gezondheid of de goede zeden, de bescherming van de goede naam of de rechten van anderen, om de verspreiding van vertrouwelijke mededelingen te voorkomen of om het gezag en de onpartijdigheid van de rechterlijke macht te waarborgen.*

Derhalve is het recht op reputatie volgens het Verdrag geen subjectief recht van het individu, maar een van de gronden waarop staten het recht

op vrijheid van meningsuiting en de persvrijheid legitiem kunnen, en niet moeten, beperken. Deze scherpe keuze van de Verdragsopstellers is door het Hof gehonoreerd in de vroege jurisprudentie over artikel 8 EVRM, waarin het werd gesteld 'that the right to honor and good name as such is not protected'[251] onder de reikwijdte van het recht op privacy. Ook in 2000 oordeelde het Hof nog in een zaak 'that the applicant's complaint relates to a perceived affront to his dignity and reputation caused by statements made by the trial judge when handing down sentence and by the Court of Appeal when upholding that sentence. This is not a matter which falls within the protection guaranteed by Article 8 of the Convention. It follows that this complaint is incompatible ratione materiae with the provisions of the Convention [].'[252]

Geleidelijk heeft het Hof echter aanvaard dat een persoon onder bepaalde omstandigheden met succes een zaak kan indienen waarin het respect voor zijn eer en goede naam centraal staat. Ten eerste heeft het Hof, via de leer van positieve verplichtingen, geoordeeld dat staten in bepaalde gevallen de vrijheid van meningsuiting dienen te beperken om respect voor iemands reputatie en eer te waarborgen.[253] Onder andere heeft het geoordeeld dat '[] where a question arises of interference with private life through publication in mass media, the State must find a proper balance between the two Convention rights involved, namely the right to respect for private life guaranteed by Article 8 (Art. 8) and the right to freedom of expression guaranteed by Article 10 (Art. 10) of the Convention.' [254]

Voortbouwend op deze lijn, waarbij het Hof stapje voor stapje bredere bescherming onder het recht op privacy erkende, heeft het Hof uiteindelijk in 2007 geoordeeld dat artikel 8 EVRM een volwaardig subjectief recht op de bescherming van de reputatie omvat. In Pfeifer tegen Oostenrijk (2007), verwees het Hof naar zijn eerdere jurisprudentie en benadrukte het 'that a person's reputation, even if that person is criticized in the context of a public debate, forms part of his or her personal identity and psychological integrity and therefore also falls within the scope of his or her "private life". Article 8 therefore applies.'[255] In latere jurisprudentie wordt deze lijn bevestigd[256] en geëxtrapoleerd naar de bescherming van eer en goede naam. In een zaak uit 2009 oordeelde het Hof: 'In more recent cases decided under Article 8 of the Convention, the Court has recognised reputation and also honour as part of the right to respect for private life. In Pfeifer, the Court held that a person's reputation, even if that person was criticised in the context of a public debate, formed part of his or her personal identity and psychological integrity and therefore also fell within the scope of his or her "private life". The same considerations must also apply to personal honour.'[257]

Dat betekent dat de materiele reikwijdte van artikel 8 EVRM sinds iets langer dan een decennium ook omvat het recht op bescherming van eer en goede naam en het recht op reputatie. Dat is onder meer van belang voor deepfakes die worden gemaakt van Bekende Nederlanders, maar kan ook relevant zijn voor gewone burgers. Ten aanzien van de materiele reikwijdte van het recht op privacy zijn nog twee additionele punten van belang.

Ten eerste heeft het ERHM meer in het algemeen benadrukt dat het recht op privacy ook omvat de bescherming van de psychologische en

lichamelijke integriteit en het recht van een persoon om zich vrijelijke te bewegen en te ontplooien, zowel in de privésfeer als in de publieke sfeer: 'the concept of "private life" is a broad term not susceptible to exhaustive definition. It covers the physical and psychological integrity of a person. It can sometimes embrace aspects of an individual's physical and social identity. Elements such as, for example, gender identification, name and sexual orientation and sexual life fall within the personal sphere protected by Article 8. Article 8 also protects a right to personal development, and the right to establish and develop relationships with other human beings and the outside world. Although no previous case has established as such any right to self-determination as being contained in Article 8 of the Convention, the Court considers that the notion of personal autonomy is an important principle underlying the interpretation of its guarantees.'[258] Hoe deze lijn zich precies verhoudt tot deepfakes zal afhangen van de omstandigheden van het geval, maar duidelijk is dat als het gaat een nepbericht waarmee iemands integriteit op het spel staat – wat vaak het geval zal zijn, wellicht met uitzondering van overduidelijke en onschuldige satire – artikel 8 EVRM van toepassing zal zijn.

Ten tweede heeft het EHRM zelfs geoordeeld dat het recht op intellectueel eigendom en het portretrecht onder omstandigheden onder artikel 8 EVRM wordt beschermd. In de zaak Bogomolova tegen Rusland (2017) gaf een vrouw toestemming voor het nemen van een foto van haar zoon in de publieke ruimte. De foto wordt zonder haar medeweten en toestemming gebruikt voor een folder over het adopteren van weeskinderen, waardoor onterecht de indruk kan ontstaan dat haar zoon wees is. De vrouw wordt in nationale aanleg in het ongelijk gesteld; het EHRM vindt dit een schending van het recht op privacy onder art. 8 EVRM. Dat doet het niet omdat er reputatieschade is ontstaan of haar eer en goede naam is geschaad of dat van haar kind, maar omdat er geen toestemming is gegeven. 'the Court has stated that a person's image constitutes one of the chief attributes of his or her personality, as it reveals the person's unique characteristics and distinguishes the person from his or her peers. The right to the protection of one's image is thus one of the essential components of personal development and presupposes the right to control the use of that image. It mainly presupposes the individual's right to control the use of that image, including the right to refuse publication thereof.'[259] Als deze uitspraak zou worden geëxtrapoleerd naar het zonder toestemming hergebruiken van beeld- of audiomateriaal zonder toestemming in het algemeen, dan kan dit grote gevolgen hebben voor de legitimiteit van deepfakes die gebruikmaken van materiaal dat van het internet is gehaald (zie verder paragraaf 3.4).

3.3.3 Vrijheid van Meningsuiting

Ook de vrijheid van meningsuiting is een recht dat door het Europees Hof voor de Rechten van de Mens een zeer ruime reikwijdte is toegekend. Daarbij zijn vijf punten van belang om te benadrukken.

Ten eerste omvat het recht op vrijheid van meningsuiting niet alleen het recht om feitelijke informatie te verspreiden en te ontvangen, het gaat daarbij ook om subjectieve kwalificaties en meningen. 'En réponse aux griefs des autres requérants selon lesquels le verdict de culpabilité aurait désigné Fırat Dink comme une cible pour les groupes ultranationalistes, qui l'ont finalement

assassiné, la Cour réitère ses considérations concernant les obligations positives de l'Etat en matière de liberté d'expression. Elle estime aussi que les obligations positives en la matière impliquent, entre autres, que les Etats sont tenus de créer, tout en établissant un système efficace de protection des auteurs ou journalistes, un environnement favorable à la participation aux débats publics de toutes les personnes concernées, leur permettant d'exprimer sans crainte leurs opinions et idées, même si celles-ci vont à l'encontre de celles défendues par les autorités officielles ou par une partie importante de l'opinion publique, voire même sont irritantes ou choquantes pour ces dernières.'[260]

Ten tweede omvat het recht ook het bedrijven van satire. Een deel van de deepfakes over politici kunnen vermoedelijk gelijk worden geschaard met de jurisprudentie over cartoons. Daarvoor geldt een grote ruimte, alhoewel het Hof ook daar grenzen aan stelt. 'Certes, cette provocation relevait de la satire dont la Cour a dit qu'il s'agissait d'une « forme d'expression artistique et de commentaire social [qui] par ses caractéristiques intrinsèques d'exagération et de distorsion de la réalité, (...) vise naturellement à provoquer et à susciter l'agitation. Elle a ajouté aussi que toute atteinte au droit d'un artiste de recourir à pareil mode d'expression doit être examinée avec une attention particulière. Toutefois, il n'en reste pas moins que le créateur, dont l'œuvre relève de l'expression politique ou militante, n'échappe pas à toute possibilité de restriction au sens du paragraphe 2 de l'article 10 : quiconque se prévaut de sa liberté d'expression assume, selon les termes de ce paragraphe, des « devoirs et responsabilités».'[261] Het kan daarbij zelfs gaan om een fictief interview. 'In this article the first applicant had wished to criticise the national hysteria after Mr Maier's accident. The essential statement behind the impugned fictitious quotation of Mr Eberharter's thought was that he had every reason to be happy about his strong rival dropping out and the consequential chance of his winning, but had not expressed this openly. In reality, Mr Eberharter had had extraordinary ski-racing successes after Mr Maier's injury. Almost everyone in Mr Eberharter's position would have had the same thought deep down inside and the statement did not imply that he had reprehensible character traits. In any event, it was clearly recognisable that he had not expressed such words at all.'[262]

Ten derde omvat het recht op vrijheid van meningsuiting ook het recht om controversiële of zelfs beledigende uitingen te doen. 'The Court's supervisory functions oblige it to pay the utmost attention to the principles characterising a "democratic society". Freedom of expression constitutes one of the essential foundations of such a society, one of the basic conditions for its progress and for the development of every man. Subject to paragraph 2 of Article 10, it is applicable not only to "information" or "ideas" that are favourably received or regarded as inoffensive or as a matter of indifference, but also to those that offend, shock or disturb the State or any sector of the population. Such are the demands of that pluralism, tolerance and broadmindedness without which there is no "democratic society". This means, amongst other things, that every "formality", "condition", "restriction" or "penalty" imposed in this sphere must be proportionate to the legitimate aim pursued.'[263]

Ten vierde kent het Hof een standaardbenadering voor wat betreft zaken waarin het recht op

privacy van de ene burger met het recht op vrijheid van meningsuiting van de andere burger of mediaorganisatie botst. De criteria die het meeweegt als het gaat om de publicatie van privéaspecten van iemands persoonlijke leven zijn: (I) Contribution to a debate of public interest, (II) The degree to which the person concerned is well known, (III) Prior conduct of the person concerned, (IV) Method of obtaining the information and its veracity concerned en (V) Content, form and consequences of the impugned article. Als het gaat om een mogelijke schending van de reputatie van een persoon dan gaat het onder meer om (I) The existence of an objective link between the impugned statement and the person claiming protection under Article 10 § 2 of the Convention, (II) The level of seriousness of the attack on reputation, (II) Benchmarks and elements in assessing whether the interference was proportionate to the legitimate aim of the protection of reputation, waaronder valt inhoudelijke elementen zoals de vorm en wijze van meningsuiting en het onderscheid tussen feitelijke mededelingen en waardeoordelen, en contexuele elementen, zoals de functie en bekendheid van de person in kwestie.[264]

Ten vijfde en tot slot is van belang dat het EHRM een flink aantal zaken over de vrijheid van meningsuiting op het internet heeft gewezen. Daarbij heeft het onder meer onderkend dat het internet bij uitstek een broedplaats voor creatieve expressie is, zoals door middel van memes en andere grappige en kunstzinnige uitingen, dat anonimiteit een belangrijk onderdeel kan zijn van het recht op vrije expressie, zeker in landen waar de overheid streng straft voor kritiek op de zittende macht, en dat grosso modo dezelfde standaarden gelden voor het online doen van uitingen als voor het offline doen van uitingen. Ook is er een aantal zaken gewezen over de verantwoordelijkheid van internet intermediairs om schadelijke uitingen tegen te gaan. Die zaken zullen kort worden aangestipt in een volgend hoofdstuk, omtrent handhaving en toezicht van materieelrechtelijke bepalingen (hoofdstuk 5).

3.3.4 Legitimate expectation

De meeste zaken waarin het Europees Hof voor de Rechten van de Mens oordeelt over een conflict tussen het recht op vrijheid van meningsuiting van de ene partij en het recht op privacy van de andere partij, kiest het een case-by-case benadering, waarin het met name kijkt naar de omstandigheden van het geval. Het kiest dan een 'balancing' benadering, waarin het de verschillende belangen van de verschillende partijen tegen elkaar afzet en afweegt. Daar zijn weinig algemene lessen uit te trekken, behalve dan redelijk voor de hand liggende punten, zoals dat als een uiting een publiek belang dient, dit zwaarder zal wegen dan als het slechts een privébelang dient, en dat hoe gevoeliger de gegevens die worden verspreid over de persoon die een beroep doet op zijn privacy, des te hoger de lat moet zijn om deze uiting legitiem te achten. Wel heeft het Hof zich in een serie arresten uitgelaten over de verhouding tussen de vrijheid van meningsuiting, vaak van de pers, en het recht op privacy van publieke figuren. Hieronder volgt een korte bespreking, aangezien deze jurisprudentie van belang kan zijn met betrekking tot deepfakes over Bekende Nederlanders.

Zeven jaar nadat het EHRM in Halford tegen het Verenigd Koninkrijk (1997) voor het eerst de doctrine van de "reasonable expectation of privacy" had aangenomen, heeft het Hof een nieuwe doctrine aangenomen, namelijk die van

de “legitimate expectation of privacy”. Het begrip “gewettigd vertrouwen” wordt voornamelijk toegepast in zaken die draaien om een conflict tussen twee particuliere partijen, waarbij de ene zich beroept op het recht op bescherming van de persoonlijke levenssfeer en de andere op de vrijheid van meningsuiting. Hoewel natuurlijke personen alleen een klacht kunnen indienen tegen een lidstaat, en niet tegen een andere burger of particuliere organisatie, kunnen zij wel opkomen tegen vonnissen die op nationaal niveau zijn gewezen en waarin een nationale rechter een conflict tussen deze twee rechten heeft geadresseerd. Om te bepalen of een inmenging in een van deze rechten (bijvoorbeeld het al dan niet door de overheid goedkeuren van een meningsuiting die tot ongenoegen van de ene of de andere partij inbreuk zou hebben gemaakt op het recht op privacy) noodzakelijk was in een democratische samenleving, zal het EHRM onder meer nagaan of de partij die zich op haar recht op privacy beroept een legitieme verwachting van privacy had. Het gebruikt dit begrip in plaats van de “reasonable expectation”, onder meer omdat het vreest dat de toepassing van een “reasonable expectation” toets ertoe zou kunnen leiden dat beroemdheden en royalty’s zich niet op het recht op privacy zouden kunnen beroepen wanneer zij structureel worden geobserveerd door en hun levens stelselmatig worden besproken in de pers. Het argument zou dan kunnen zijn dat zij redelijkerwijs niet kunnen verwachten dat zij nog privacy hebben in de publieke ruimte.

De eerste keer dat het leerstuk van de legitieme verwachting van privacy werd gebruikt, was in Von Hannover tegen Duitsland (2004).[265] Caroline Von Hannover, de prinses van Monaco, had meer dan tien jaar zonder succes voor Duitse rechtbanken geprocedeerd. Zij stelde dat zij, zodra zij haar huis verliet, voortdurend werd achtervolgd door paparazzi die elke beweging van haar volgden, of het nu ging om het oversteken van de weg, het ophalen van haar kinderen van school, het doen van boodschappen, het maken van een wandeling, het beoefenen van sport of het op vakantie gaan. Het Hof erkent dat er een inbreuk is geweest op haar privéleven. Het Hof benadrukt dat zelfs publieke personen een ‘gewettigd vertrouwen’ hebben dat zij privacy kunnen genieten in de privésfeer. Het Hof beklemtoonde dat een grotere waakzaamheid bij de bescherming van de persoonlijke levenssfeer geboden is, onder meer om het hoofd te kunnen bieden aan de nieuwe communicatietechnologieën die het mogelijk maken persoonsgegevens op te slaan en te reproduceren, en voegde daaraan toe dat ‘the distinction drawn between figures of contemporary society *“par excellence”* and “relatively” public figures has to be clear and obvious so that, in a State governed by the rule of law, the individual has precise indications as to the behaviour he or she should adopt. Above all, they need to know exactly when and where they are in a protected sphere or, on the contrary, in a sphere in which they must expect interference from others, especially the tabloid press.’ [266]

In Standard Verlags GMBH v. Oostenrijk (nr. 2) (2009),[267] merkte het Hof op dat een artikel waarover in nationale aanleg een oordeel was geveld, betrekking had op geruchten over het privé- en gezinsleven van politici en benadrukte dat ook politici een “legitieme verwachting van privacy” hebben. Het maakte een onderscheid tussen de vermeende huwelijksproblemen van een politicus en zijn gezondheidstoestand, die weliswaar tot de persoonlijke sfeer behoort,

maar wel van invloed kan zijn op de uitoefening van zijn functies. Omdat het privéleven van het presidentieel echtpaar geen rol had gespeeld tijdens de tweede ambtstermijn van de president en omdat de vermeende buitenechtelijke relatie van de First Lady geen enkel verband hield met de publieke functies en verantwoordelijkheden van de president, oordeelde het Hof dat de publicatie niet bijdroeg tot enig publiek debat ten aanzien waarvan de pers haar rol van "publieke waakhond" dient te vervullen, maar enkel diende om de nieuwsgierigheid van een bepaald lezerspubliek te bevredigen. Het feit dat de vermeende buitenechtelijke affaire van de First Lady met de leider van een extreemrechtse of neofascistische partij (FPO) was, deed daar niets aan af. Het EHRM oordeelde dus dat de verplichting van de kranten om schadevergoeding te betalen aan de eisers, zoals was bevolen door de rechters in nationale aanleg, niet onevenredig was aan het legitieme doel van bescherming van de privacy van publieke personen.

Alhoewel er ook zaken zijn waarin het EHRM juist voorrang geeft aan de vrijheid van meningsuiting boven het recht op privacy van publieke personen en waarin het benadrukt dat het feit dat personen zelf, actief de publiciteit hebben opgezocht dient te worden meegewogen in de mate waarin iemand nog een beroep kan doen op zijn legitieme privacyverwachting, kiest het vaak een privacyvriendelijk pad. In Ruusunen v. Finland (2014) accepteerde het Hof bijvoorbeeld zelfs een strafrechtelijke veroordeling op nationaal niveau van een voormalige vriendin van de minister-president die een autobiografisch boek had geschreven over haar relatie. Het EHRM benadrukte dat de in het boek uiteengezette feiten niet ter discussie stonden en op een meelevende manier werden gepresenteerd en dat de stijl niet provocerend of overdreven was. De premier was duidelijk een publiek figuur en hij had zelfs toestemming gegeven om zijn foto op de omslag van het boek te gebruiken.[270] Het Hof benadrukte dat, ook al lag de nadruk in het boek op het privéleven van de klager, het boek niettemin elementen van openbaar belang bevatte en dat het merendeel van de informatie over het privéleven van de premier reeds op ruime schaal openbaar was gemaakt. Ondanks al deze overwegingen achtte het EHRM de strafrechtelijke veroordeling redelijk, omdat sommige informatie in het boek betrekking had op zijn seksuele en intieme leven, die nog niet openbaar waren gemaakt. Het Hof volgde een vergelijkbare benadering in Ojala en Etukeno Oy v. Finland (2014) en Salumaki v. Finland (2014).[271]

In Alhpa Doryforiki Tileorasi Anonymi Etairia v. Griekenland (2018),[272] had een omroep een televisieprogramma met de naam "Jungle" vertoond waarin drie video's werden uitgezonden die met een verborgen camera waren opgenomen. Op de eerste video was te zien hoe A.C., destijds lid van het Griekse parlement en voorzitter van het interparlementair comité voor elektronische kansspelen, een speelhal binnenging en op twee automaten speelde. De tweede video toonde een ontmoeting tussen A.C. en medewerkers van de televisiepresentator van "Jungle", M.T., tijdens welke de eerste video werd getoond aan A.C. De derde video toonde een ontmoeting tussen A.C. en M.T. in het kantoor van laatstgenoemde. Hoewel het EHRM opmerkte dat het verslag in kwestie niet zonder politiek belang was, vond het desalniettemin een privacyschending omdat de parlementariër niet op de hoogte was gesteld van het feit dat hij werd gefilmd en het EHRM er niet van overtuigd was dat het heimelijk filmen noodzakelijk was voor de tweede en derde video.

Derhalve oordeelde het Hof dat de sanctie van € 100.000 niet alleen in overeenstemming was met het Verdrag, maar zelfs mild was, gezien de legitieme verwachting van privacy van A.C. Het Hof ging dus een stap verder dan in Standard Verlags GMBH, omdat in de Griekse zaak de publicatie direct verband hield met het publieke ambt dat de parlementariër bekleedde.

Een laatste voorbeeld is van belang omdat het daarin ging om het hergebruik van informatie die reeds publiek was gemaakt, of nog preciezer, het verwijzen naar informatie die reeds publiek was gemaakt. In Khadija Ismayilova v. Azerbaijan (nr. 3) (2020) lekte een video van de verzoekster, een bekende onderzoeksjournalist, waarin zij te zien was terwijl zij seks had met haar toenmalige vriend, uit naar de pers. Een krant publiceerde het materiaal niet, maar verwees ernaar en leverde op negatieve wijze commentaar. De regering betoogde dat verzoekster geen gewettigd vertrouwen in de bescherming van haar persoonlijke levenssfeer had, niet alleen omdat zij een publiek figuur was, maar ook omdat het materiaal waarnaar in de reportage werd verwezen reeds door andere media openbaar was gemaakt. Het Hof was het daar niet mee eens. 'It is true that, once a person's privacy has been breached and the information about it has entered into public domain, the damage is already done and it is virtually impossible to restore the situation to when the breach had never happened. However, while responsible reporting on matters of public interest in accordance with the ethics of journalism is protected by the Convention, there can be no legitimate public interest in exploiting an existing breach of a person's privacy for the purpose of satisfying the prurient curiosity of a certain readership, publicly ridiculing the victim and causing them further harm.'[273] Zelfs wanneer gevoelige informatie al op grote schaal op internet circuleert en in verschillende media wordt geciteerd en zelfs wanneer een specifiek medium die informatie niet zelf publiceert, maar er alleen naar verwijst, kan dit nog steeds worden beschouwd als een schending van de legitieme verwachting van privacy van een publieke figuur, aldus het Europese Hof voor de Rechten van de Mens.

Toch is er evenzoveel, zo niet meer jurisprudentie van het EHRM waarin de balans negatief uitvalt voor publieke figuren en het Hof stelt dat zij een grotere inmenging op hun privésfeer dienen te dulden dan gewone burgers.[274] Uit het voorgaande blijkt derhalve dat het Hof zowel een zeer brede reikwijdte toekent aan het recht op privacy, waaronder valt de bescherming van reputatie, eer en goede naam, als aan het recht op vrijheid van meningsuiting. De vraag of er in een geval een schending van het ene of het andere recht is geweest zal het Hof altijd beantwoorden in de concrete omstandigheden van het geval, waarbij het rekening houdt met tal van factoren, zoals de inbreukmakendheid op de privésfeer van de uiting, het belang dat met de uiting is gemoeid, de vraag of de persoon in kwestie een publiek persoon is en of de persoon zelf zaken openbaar heeft gemaakt. Voor deepfakes volgt dat als het komt tot een botsing van deze twee rechten, er slecht in algemene zin iets kan worden gezegd over welk recht de doorslag zal geven. Wel kan de wetgever ervoor kiezen om hier sturing in aan te brengen.

3.3.5 Conclusie

In deze paragraaf is stilgestaan bij de inkadering van deepfakes door het Europees Verdrag voor de Rechten van de Mens en dan met name 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, goede naam en reputatie. Ook heeft het Hof bepaald 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, beoordeeld, kan niet in algemene zin worden gezegd hoe deze twee rechten zich bij deepfaketoepassingen tot elkaar verhouden. Dit zal per zaak moeten worden bekeken. Wel zijn twee bijzondere punten van belang. Enerzijds heeft het Hof een uitgebreide doctrine aangaande wat het noemt de 'reasonable expectation of privacy'; daaruit volgt dat mensen ook in een werkomgeving of in de publieke ruimte mogen verwachten dat hun privacy wordt beschermd. Zelfs als een persoon extreme seksuele afbeeldingen van zichzelf op het internet zet, dan nog mag hij verwachten dat zijn privacy door anderen wordt gerespecteerd. Dat is van belang omdat hieruit volgt dat het recht op privacy op veruit het meeste materiaal dat wordt gebruik voor het genereren van deepfakes van toepassing zal zijn. Anderzijds geldt er een speciale doctrine voor bekende personen. Dit is van belang omdat de meeste deepfakes worden gemaakt ofwel over personen in de directe omgeving van de maker ofwel over bekende personen. Deze bekende personen moeten, zeker als het politici of ambtsdragers betreft, volgens het Hof meer dulden in termen van inperkingen in hun privésfeer en hun reputatie, eer en goede naam dan gewone burgers. Toch heeft het Hof eveneens benadrukt dat dergelijke inperkingen alsnog proportioneel dienen te zijn en dat ook publieke personen een recht op privacy toekomt.

3.4 Portretrecht

Het is van belang het intellectueel eigendomsrecht te bestuderen, omdat dit rechtsgebied bescherming biedt aan de maker van een werk dat wordt gebruikt voor de vervaardiging van gemanipuleerd materiaal, ruimte biedt aan het hergebruik van materiaal onder bepaalde voorwaarden, bescherming biedt aan de geportretteerde en regels bevat aangaande het downloaden en gebruiken van bestaand of gemanipuleerd materiaal. In hoeverre kan een persoon gebruik maken van het portretrecht bij deepfakes? Welke mate van gelijkenis met de oorspronkelijke beeltenis van een persoon moet er zijn om deze doctrine van toepassing te laten zijn? Deze kwestie zal worden toegelicht in deze paragraaf.

3.4.1 EVRM en Auteurswet

Het portretrecht is onderdeel van het auteursrecht en is te vinden in de artikelen 19 – 21 van de Auteurswet. Dit recht is verbonden met artikel 8 van het EVRM, het recht op bescherming van het persoonlijke leven. Dit komt onder meer naar voren in arresten van het EHRM, zoals de eerder besproken zaak van Bogomolova en de zaak van Reklos en Davourlis v. Griekenland (2009),[275] waarin het Hof oordeelde dat het fotograferen van een persoon in de privésfeer zonder diens toestemming, zoals in deze zaak, een schending

van de privacy oplevert. Het recht op privacy houdt ook het recht op een eigen identiteit en zelfontplooiing in. Tevens is iemands portret een van de belangrijkste gegevens om iemands persoonlijke kenmerken te onderscheiden, aangezien het de unieke kenmerken van deze persoon onthult en de persoon zich zo onderscheidt van zijn of haar leeftijdsgenoten. Onderstaande evaluatie zal met name in gaan op de situatie onder het Nederlands recht en dan specifiek de Auteurswet.

Binnen het Nederlands portretrecht kan er een onderscheid worden gemaakt tussen het portret dat in opdracht is gemaakt of ten behoeve van de geportretteerde is gemaakt, wat onder de artikelen 19 en 20 Auteurswet valt, en het portret zonder opdracht, wat in artikel 21 Auteurswet is geregeld. Binnen het kader van het onderzoek over deepfakes is met name die laatste variant van belang. Artikel 21 van de Auteurswet luidt:

> *"Is een portret vervaardigd zonder daartoe strekkende opdracht, den maker door of vanwege den geportretteerde, of te diens behoeve, gegeven, dan is openbaarmaking daarvan door dengene, wien het auteursrecht daarop toekomt, niet geoorloofd, voor zoover een redelijk belang van den geportretteerde of, na zijn overlijden, van een zijner nabestaanden zich tegen de openbaarmaking verzet."*

Het artikel spreekt uitdrukkelijk over het openbaar maken van een portret, dit betekent dat deze bepaling beschermt tegen de openbaarmaking van een portret, maar niet tegen het maken, bewaren, opslaan en ander gebruik van portretten.[276] Een geportretteerde kan daar uiteraard wel tegen optreden, middels een procedure op grond van artikel 6:162 BW. Waar in het artikel wordt gesproken van 'een redelijk belang' moet worden uitgegaan van een belangenafweging, zo heeft de Hoge Raad in het arrest Ferdi E. geoordeeld.[277] In dit arrest heeft de Hoge Raad een belangenafweging gemaakt tussen enerzijds het recht op eerbiediging van de persoonlijke levenssfeer van Ferdi E. (een bekende misdadiger), tegenover het recht op vrijheid van meningsuiting aan de kant van de uitgeverij. Het recht op eerbiediging van de persoonlijke levenssfeer kent geen 'absoluut gewicht' dat in beginsel groter is dan dat van het recht op vrijheid van meningsuiting. De Hoge Raad beslist dat beide vrijheden, enerzijds de ontplooiing van het individu en anderzijds voor een democratische samenleving, even belangrijk zijn, zodat er geen rangorde tussen deze twee vrijheden bestaat. [278]

Als een geportretteerde zich tegen openbaarmaking van het portret wil verzetten, moet hij een redelijk belang kunnen aantonen. Volgens de Memorie van Toelichting bij artikel 21 Auteurswet, wordt hieronder verstaan 'een afbeelding van het gelaat van een persoon, met of zonder die van verdere lichaamsdelen, op welke wijze zij ook vervaardigd is'.[279] Hierbij is van belang dat het voor dit artikel niet nodig is dat de geportretteerde poseerde voor de foto. De Hoge Raad heeft geoordeeld dat het niet noodzakelijk is dat de oogpartij zichtbaar is, een kenmerkende lichaamshouding kan al bijdragen aan de herkenbaarheid van het gelaat. Ook is herkenbaarheid voor enkelingen al voldoende.[280] Daarnaast hoeft het geheel of gedeeltelijk onherkenbaar maken van het gelaat van de geportretteerde er niet aan af te doen dat het gaat om een portret in de zin van artikel 21 Auteurswet, omdat de identiteit van een persoon uit overige zaken op de afbeelding kan blijken.[281]

In geval van een portret dat niet in opdracht is vervaardigd, is openbaarmaking niet toegestaan voor zover een redelijk belang zich daartegen verzet. Dit is zowel gericht tegen de auteursrechthebbende als tegen derden.[282] In eerste instantie wordt onder een redelijk belang het zedelijke belang beschouwd. Deze categorie ziet op de eerbiediging van de persoonlijke levenssfeer, zo heeft de Hoge Raad in 1988 geoordeeld dat een redelijk belang gelinkt is aan het recht op privacy van artikel 8 EVRM.[283] Daarvoor werden commerciële belangen door de Hoge Raad erkend als redelijke belangen. De Hoge Raad overwoog in dit arrest dat 'van een redelijk belang ook sprake [kan] zijn, wanneer de populariteit van geportretteerden, verworven in de uitoefening van hun beroep, van dien aard is, dat een commerciële exploitatie van die populariteit door enigerlei wijze van openbaarmaking van hun portretten mogelijk wordt. Het belang van de geportretteerden om dan in de voordelen van zulke exploitatie mee te kunnen delen door de openbaarmaking van hun portretten voor commerciële doeleinden niet te hoeven toelaten zonder daarvoor vergoeding te ontvangen, is een redelijk belang in de zin van art. 21.'[284]

Dat houdt in dat als een persoon over verzilverbare populariteit beschikt, hij zich tegen publicatie van zijn portret kan verzetten. Voornamelijk bekendheden zullen zich hierop beroepen. Het houdt als ware in dat bekendheden kunnen optreden tegen het gebruik van hun portret zonder hun toestemming en een vergoeding kunnen eisen, gebaseerd op het bedrag dat zij hadden kunnen verdienen als zij hadden ingestemd met het gebruik van het portret. Geportretteerden hoeven dus niet te accepteren dat hun portret voor commerciële doelen wordt gebruikt, zonder dat zij daarvoor een redelijke vergoeding ontvangen. Dit volgt uit het arrest Cruijf/Tirion, waarin de Hoge Raad bovendien over personen zonder publieke bekendheid zei "dat zij openbaarmaking van hun portret in beginsel niet behoeven te dulden".[285] Hiervoor zegt Hugenholtz dat onbekende Nederlanders zodoende een sterk portretrecht genieten, ook in gevallen dat de privacy van deze persoon niet in het geding is.[286]

Hieronder zullen een aantal zaken kort worden aangestipt die van belang zijn bij deepfakes, namelijk de jurisprudentie over lookalikes, over de samenvoeging van twee of meerdere portretten, de rechten van nabestaanden, pornografische deepfakes, rechten op audiobestanden, morele rechten en de privékopie-exceptie. Daarbij moet worden bedacht dat niet alleen op het materiaal dat wordt gebruikt voor het genereren van een deepfake auteursrecht kan liggen, maar vervolgens ook op de gegenereerde deepfake als dat als een originele schepping wordt gezien.

3.4.2 Lookalike

In het geval van een lookalike is het de bedoeling van degene die het portret maakt om bij de afbeelding van de ene persoon een beeld van een ander persoon op te roepen.[287]

Op 24 juni 2005 deed de rechtbank uitspraak in een geschil tussen de Gouden Gids en Yellow Bear.[288] Indertijd stond Katja Schuurman model voor advertenties van de Gouden Gids. Vervolgens heeft Yellow Bear met een van achteren afgebeelde lookalike van Katja Schuurman een vergelijkbare reclame als die van de Gouden Gids gemaakt. De vrouw op de reclame van Yellow Bear had een vergelijkend pak aan, eenzelfde houding, met hetzelfde soort haar, maar van de achterkant gefotografeerd waardoor het gezicht

niet herkenbaar in beeld kwam. De rechtbank oordeelde dat ook deze uitsluitend van achteren afgebeelde lookalike een portret is in de zin van artikel 21 Auteurswet was en dat de reclame afbreuk deed aan het persoonlijke eigen recht van Katja Schuurman om zelf te bepalen op welke wijze en voor wie zij reclame wenste te maken.[289] Daarnaast oordeelde de rechtbank dat er geen sprake was van een parodie omdat commerciële concurrentiemotieven een doorslaggevende rol hadden gespeeld.[290]

In een andere zaak, waar het ook ging om een bekend persoon, werd anders geoordeeld. Op 2 juni 2020 heeft het Hof Amsterdam uitspraak gedaan in een zaak tussen Max Verstappen en Picnic.[291] Verstappen werd eerder het gezicht van een van de reclames van supermarkt Jumbo. Als reactie hierop besloot online supermarkt Picnic in 2016 een lookalike van Verstappen te gebruiken voor een reclame, waarbij de lookalike dezelfde outfit aanhad en waarbij deze persoon de diensten van Picnic aanprees. Het filmpje heeft in totaal een dag online gestaan en is verwijderd nadat Verstappen hierom vroeg. Verstappen is vervolgens naar de rechter gegaan en kreeg van de rechter gelijk. Picnic had het portretrecht van Verstappen geschonden, door het gebruik van een lookalike. Picnic moest Verstappen een schadevergoeding betalen. Verstappen was het niet eens met de hoogte van de schadevergoeding, maar ook Picnic was het er niet mee eens en ging in hoger beroep. Het Hof oordeelde dat Picnic geen inbreuk op het portretrecht van Verstappen heeft gemaakt, omdat het voor de aanschouwer duidelijk is dat het niet Verstappen betreft maar dat het ging om een persiflage van het optreden van Verstappen in de reclame van Jumbo. Het Hof oordeelde dat het duidelijk was dat het een lookalike betrof en het portretrecht niet reikte tot bepaalde kenmerken van iemand die door een ander persoon zijn uitgebeeld. Deze redenering staat lijnrecht tegenover de uitspraak van de Hoge Raad in de Katja Schuurman/Yellow Bear-zaak. Verstappen kon – omdat het volgens het Hof geen portet betrof – geen aanspraak maken op de bescherming van artikel 21 Auteurswet. Bovendien oordeelde het Hof dat Picnic niet onrechtmatig handelde en het filmpje niet van zodanige aard was dat het de eer en goede naam heeft aangetast of dat zijn zakelijke belangen zijn geschaad. Wel is van belang dat er in de literatuur redelijk wat kritiek is op het criterium dat Hof aanlegt.

In beide zaken komt het aspect parodie naar voren, waarbij het in het ene arrest niet wordt aangenomen en in het andere wel. In beide zaken wordt vastgesteld dat de bedoeling was dat het publiek de bekende Nederlander zou herkennen in de parodie. Echter, het verschil tussen deze zaken is dat in het geval van de parodie op Katja Schuurman het publiek zou denken dat het daadwerkelijk Katja Schuurman betrof terwijl de rechter in het geval van de parodie op Max Verstappen oordeelde dat het publiek zou beseffen dat het om een lookalike gaat.[292] Het gehanteerde criterium door het Hof wordt (nog) niet gehanteerd in de rechtspraak door de Hoge Raad of het EHRM.

Duidelijk is dat het parodie-aspect een belangrijke factor is in het al dan niet toewijzen van een vordering op grond van artikel 21 Auteurswet, in ieder geval in zaken die bekendheden betreffen. Dat het betwistbaar is of een afbeelding van een lookalike het portretrecht van de gelijkende persoon schendt, is relevant in het kader van deepfakes. De beelden en/of audiofragmenten zijn vaak niet 'echt' een representatie van wat

iemand heeft gezegd en/of gedaan, maar lijken daarop. Daarom zullen rechters wellicht naar analogie van de jurisprudentie omtrent lookalikes oordelen of geschillen waarop een beroep wordt gedaan op het portretrecht. Daarbij geldt wel dat er lang niet altijd commerciële motieven zijn en is ook de vraag waar de grens tussen parodie en ernst ligt in de deepfake context. Een doorslaggevend element in de zaak Verstappen was nu juist dat de goede verstaander snapte dat het een knipoog betrof, terwijl een doel van veel deepfakes juist is om echt en waarachtig over te komen.

3.4.3 Samenvoeging van twee gezichten tot een foto

Door middel van deepfakes kunnen ook nieuwe gezichten worden gecreerd uit twee of meerdere bestaande gezichten. Daarnaast bieden social media platforms/apps gebruikers de mogelijkheid om samen met vrienden gebruik te maken van een 'Faceswap'. Hierbij worden kenmerken van de gezichten van beide personen samengevoegd tot een nieuw gezicht, zodat iemand kan zien hoe hij er uit ziet met bijvoorbeeld de neus of ogen van een vriend. De vraag is of het portretrecht in dergelijke gevallen van toepassing is en zo ja, van wie dan. Zoals eerder opgemerkt heeft de Hoge Raad geoordeeld dat het niet noodzakelijk is dat de oogpartij zichtbaar is om van iemands portret te spreken; een kenmerkende lichaamshouding kan al bijdragen aan de herkenbaarheid van het gelaat.[293] Ook bij collagefoto's geldt normaal het portretrecht. 'Ook als je een foto bewerkt, bijvoorbeeld door deze op te nemen in een collage, berust het auteursrecht bij de maker. Je mag het beeld dus niet op die manier (in een collage) zonder toestemming verveelvoudigen of reproduceren.'[294] Dat zou kunnen betekenen dat een of meerdere personen portretrecht kunnen claimen over samengestelde foto's. Veel apps worden gebruikt met toestemming van de geportretteerden, immers zijn zij vaak zelf de maker van de Faceswap, waardoor artikelen 19-20 Auteurswet van toepassing zijn. In gevallen waarbij er geen sprake is van een opdracht, is artikel 21 Auteurswet van toepassing. Dan kunnen dergelijke collages op bezwaren stuiten.

3.4.4 Nabestaanden

Het portretrecht komt de geportretteerde toe en, na diens overlijden, aan een nabestaande. Onder nabestaanden worden de ouders, echtgenoot, geregistreerd partner en kinderen beschouwd, andere erfgenamen kunnen niet de portretrechten van de overledene verkrijgen. Deze nabestaanden kunnen gedurende tien jaar na diens dood het portretrecht inroepen, dit is bepaald in artikel 25a Auteurswet. Maar net als de geportretteerde zelf, kunnen de nabestaanden zich ook op het portretrecht beroepen na deze tien jaar, mits zij een redelijk belang – dit kan zowel een privacy als een commercieel belang zijn, of met andere worden een moreel dan wel economisch belang – hebben tegen publicatie van het portretrecht. Dit is echter geen absoluut recht. Voor nabestaanden gelden dus – na het overlijden van de geportretteerde – dezelfde vereisten als voor de geportretteerde zelf.[295]

In 2018 – ongeveer twee jaar na het overlijden van Johan Cruijf – hebben zijn nabestaanden de publicatie van een boek met als omslag een foto van Cruijf geprobeerd tegen te houden. Er was sprake van een openbaarmaking van een niet in opdracht gemaakt portret, die niet in een privésetting was gemaakt. De uitgeverij heeft de nabestaanden voor het gebruik van de foto geen redelijke vergoeding geboden, terwijl dit in het geval van de verzilverbare populariteit, wel van

belang kan zijn. Uiteindelijk heeft de uitgeverij 10% van de netto omzet aangeboden als vergoeding. Dat dit niet op voorhand was aangeboden vond de rechter onvoldoende om de publicatie te kunnen verbieden. De voorzieningenrechter oordeelde dat de afbreuk van of schade aan de reputatie van Cruijf onvoldoende aannemelijk was gemaakt, waardoor het publicatieverbod werd afgewezen.

De vraag in het kader van deepfakes is hoe rechters zullen oordelen over het gebruik van afbeeldingen van overleden personen. Daarbij is van belang in welke gevallen nabestaanden bezwaar kunnen maken tegen het gebruik van het portret van hun dierbare en zullen er op termijn conflicten ontstaan tussen nabestaanden (vrouw wil haar overleden man laten spreken op zijn begrafenis als deepfake, maar de kinderen zijn daarop tegen). Rechters zullen hier kijken naar het redelijk belang van de diverse partijen en daarin een belangenafweging moeten maken per geval.

3.4.5 Pornografische deepfake

Gelijkend aan een pornografische deepfake is het fenomeen 'wraakporno'. Het verschil hierbij is wel dat in het geval van wraakporno de geportretteerde daadwerkelijk ook de persoon is die de maker van het portret wil portretteren. Dit is een klein onderscheid; desalniettemin kan de jurisprudentie over dit onderwerp op het gebied van het portretrecht waardevol zijn voor mogelijke zaken betreffende pornografische deepfakes.

In 2014 heeft toenmalig minister van Veiligheid en Justitie Opstelten naar aanleiding van Kamervragen opheldering verschaft omtrent dit onderwerp.[296] De Nederlandse wetgeving zou in beginsel geen lacunes bevatten om op te kunnen treden tegen het ongewenst verspreiden van seksueel getinte beelden.[297] Zowel het strafrecht als het civiel recht voorzien in een aantal bepalingen om wraakporno te kunnen vervolgen. Een slachtoffer dat herkenbaar in beeld is gebracht op ongewenste beelden, kan een beroep doen op artikel 19 tot en met 21 Auteurswet, stelde de minister. Op basis van vaste rechtspraak van de Hoge Raad, is in het geval van een zaak over naaktbeelden, snel sprake van een redelijk belang aan de kant van de geportretteerde. Zo heeft de Hoge Raad in het naturistengids-arrest geoordeeld dat het enkele feit dat geportretteerde naakt is afgebeeld een redelijk belang oplevert dat zich tegen openbaarmaking verzet.[298]

In het Discodanser-arrest ging het om een beroeps-discodanser die tegen betaling een dansoptreden verzorgde in een discotheek. Tijdens deze avond werden foto's gemaakt, waarop de danser met ontbloot bovenlijf te zien was. Deze discotheek richtte zich niet in geringe mate op een homopubliek. Later heeft de exploitant van de discotheek de foto van de danser gebruikt voor in een reclamefolder, waarop onder de foto het opschrift 'Mooi bloot gezocht voor de Toplessparty'. Daarnaast is de foto ook afgedrukt op de achterzijde van de Gay-krant. De geportretteerde was niet op de hoogte van deze publicatie en maakte bezwaar tegen de publicatie van zijn portret en deed alsmede een beroep op artikel 21 Auteurswet. De Hoge Raad overwoog dat hoewel de geportretteerde medewerking had verleend aan het voortbrengen van het product of het verrichten van de dienst, dat niet met zich bracht dat de openbaarmaking van het portret in een reclame-uiting niet kon worden beschouwd als een inbreuk op de persoonlijke levenssfeer van de geportretteerde, of dat geen redelijk belang van de geportretteerde

zich meer tegen de inbreuk kon verzetten. Van inbreuk op de persoonlijke levenssfeer is in het bijzonder sprake, als de reclame-uitingen een persoon portretteren in 'een openbare sfeer van erotiek en vrijheid van opvattingen'.[299] Hoewel het bij een pornografische deepfake niet gaat om een reclame-uiting, heeft het wel alle schijn van een openbare sfeer van erotiek. Bovendien kan er ook gesproken worden van een inbreuk op de persoonlijke levenssfeer door de sfeer van erotiek.

In een zaak gewezen door rechtbank Rotterdam ging het om een bruidspaar dat koos voor een pikante fotoshoot, waarbij het bruidspaar in lingerie poseerde. Na de bruiloft vroeg de fotograaf toestemming om de foto's te publiceren op zijn website. De vrouw heeft, omdat haar man liever niet wilde dat zij herkenbaar in beeld zou komen, enkele foto's geselecteerd waarop zij niet herkenbaar in beeld was en aangegeven dat deze foto's op zijn site getoond mochten worden. Twee jaar later constateerde de vrouw dat de fotograaf foto's had gebruikt waarop zij herkenbaar in beeld. De rechtbank oordeelde dat er sprake was van een onrechtmatige daad van de fotograaf en veroordeelde de fotograaf tot het betalen van een schadevergoeding. Omdat het om extra gevoelige beelden ging had de fotograaf extra zorgvuldigheid moeten betrachten. Bij het bepalen van de hoogte van de schadevergoeding onderstreepte de rechtbank dat de foto's niet alleen op de website van de fotograaf, maar ook op social media waren gezet, waardoor de foto's voor iedereen toegankelijk waren.[300]

3.4.6 Audio-opname

In hoofdstuk 2 werd onder meer ingegaan op een deepfake-audioclips van Jay-Z. De vraag rijst of zo'n geval ook onder het portretrecht valt. In 2012 oordeelde de rechtbank Amsterdam – tevens bevestigd door het gerechtshof – dat een stemopname niet onder het bereik van het portretrecht valt. Ook als de stem informatie geeft op grond waarvan de identiteit van de persoon achter de stem eventueel kan worden vastgesteld.[301] Dit heeft de rechtbank Midden-Nederland in 2020 bevestigd. Een vrouw had een zaak aangespannen tegen Ali B en zijn management. Een aantal jaar voor de zaak was de vrouw het slachtoffer geworden van een gewapende overval op haar winkel. Na de overval werkte de vrouw mee met het tv-programma "Opsporing verzocht", waarin zij vertelde dat één van de overvallers een Ali B-accent had. Op een later moment had Ali B dit geluidsfragment in zijn theatershow gebruikt. Volgens de vrouw maakte dit inbreuk op haar portretrecht en haar recht op privacy. De rechtbank oordeelde dat een stemopname niet onder het portretrecht valt.[302] Wel zou dit in bepaalde omstandigheden onder de beschermingsreikwijdte van Artikel 8 EVRM kunnen vallen.

Morele rechten

Van belang is ook dat makers van een portretrechten hebben, omdat dit moet worden meegenomen in de uiteindelijke keuze tussen de diverse reguleringsopties (hoofdstuk 8). Artikel 25 AW bepaalt bijvoorbeeld:

> *1. De maker van een werk heeft, zelfs nadat hij zijn auteursrecht heeft overgedragen, de volgende rechten:*
> *a. het recht zich te verzetten tegen openbaarmaking van het werk zonder vermelding van zijn naam of andere aanduiding als maker, tenzij het verzet zou zijn in strijd met de redelijkheid;*
> *b. het recht zich te verzetten tegen de*

openbaarmaking van het werk onder een andere naam dan de zijne, alsmede tegen het aanbrengen van enige wijziging in de benaming van het werk of in de aanduiding van de maker, voor zover deze op of in het werk voorkomen, dan wel in verband daarmede zijn openbaar gemaakt;

c. het recht zich te verzetten tegen elke andere wijziging in het werk, tenzij deze wijziging van zodanige aard is, dat het verzet zou zijn in strijd met de redelijkheid;

d. het recht zich te verzetten tegen elke misvorming, verminking of andere aantasting van het werk, welke nadeel zou kunnen toebrengen aan de eer of de naam van de maker of aan zijn waarde in deze hoedanigheid.

2. De in het eerste lid genoemde rechten komen, na het overlijden van de maker tot aan het vervallen van het auteursrecht, toe aan de door de maker bij uiterste wilsbeschikking aangewezene.

3. Van het recht, in het eerste lid, onder a genoemd kan afstand worden gedaan. Van de rechten onder b en c genoemd kan afstand worden gedaan voor zover het wijzigingen in het werk of in de benaming daarvan betreft.

4. Heeft de maker van het werk het auteursrecht overgedragen dan blijft hij bevoegd in het werk zodanige wijzigingen aan te brengen als hem naar de regels van het maatschappelijk verkeer te goeder trouw vrijstaan. Zolang het auteursrecht voortduurt komt gelijke bevoegdheid toe aan de door de maker bij uiterste wilsbeschikking aangewezene, als redelijkerwijs aannemelijk is, dat ook de maker die wijzigingen zou hebben goedgekeurd.

Deepfakes zijn bijna per definitie aanpassingen aan bestaande werken. Vaak berust het portretrecht op de afbeeldingen niet bij de maker van de oorspronkelijke portretten die zijn gemaakt. De oorspronkelijke auteur mag zich derhalve in principe altijd verzetten tegen een deepfake die op basis van zijn werk is vervaardigd, tenzij dit 'in strijd zou zijn met de redelijkheid'. Bij parodieën of onschuldige aanpassingen zal hier minder snel sprake van zijn dan als van een normaal portret een deepfake-pornofilm wordt gemaakt.

3.4.7 Privékopie-exceptie

Tot slot is van belang de zogenoemde privékopie-exceptie uit Artikel 16c AW. Dit vormt een uitzondering op het exploitatierecht van de auteur. In principe is toestemming nodig van de auteur voor de vermenigvuldiging of verspreiding van het werk, tenzij een uitzondering van toepassing is, zoals de privékopie-exceptie. Het voorgenoemde artikel leest:

1. Als inbreuk op het auteursrecht op een werk van letterkunde, wetenschap of kunst wordt niet beschouwd het reproduceren van het werk of een gedeelte ervan op een voorwerp dat bestemd is om een werk ten gehore te brengen, te vertonen of weer te geven, mits het reproduceren geschiedt zonder direct of indirect commercieel oogmerk en uitsluitend dient tot eigen oefening, studie of gebruik van de natuurlijke persoon die de reproductie vervaardigt.

2. Voor het reproduceren, bedoeld in het eerste lid, is ten behoeve van de maker of diens rechtverkrijgenden een billijke vergoeding verschuldigd. De verplichting tot betaling van de vergoeding rust op de fabrikant of de importeur van de voorwerpen, bedoeld in het eerste lid.

3. Voor de fabrikant ontstaat de verplichting tot betaling van de vergoeding op het tijdstip dat de door hem vervaardigde voorwerpen in het verkeer

kunnen worden gebracht. Voor de importeur ontstaat deze verplichting op het tijdstip van invoer.

4. De verplichting tot betaling van de vergoeding vervalt indien de ingevolge het tweede lid betalingsplichtige een voorwerp als bedoeld in het eerste lid uitvoert.

5. De vergoeding is slechts eenmaal per voorwerp verschuldigd.

6. Bij algemene maatregel van bestuur kunnen nadere regelen worden gegeven met betrekking tot de voorwerpen ten aanzien waarvan de vergoeding, bedoeld in het tweede lid, verschuldigd is. Bij algemene maatregel van bestuur kunnen voorts nadere regelen worden gegeven en voorwaarden worden gesteld ter uitvoering van het bepaalde in dit artikel met betrekking tot de hoogte, verschuldigdheid en vorm van de billijke vergoeding.

7. Indien een ingevolge dit artikel toegelaten reproductie heeft plaatsgevonden, mogen voorwerpen als bedoeld in het eerste lid niet zonder toestemming van de maker of zijn rechtverkrijgenden aan derden worden afgegeven, tenzij de afgifte geschiedt ten behoeve van een rechterlijke of bestuurlijke procedure.

8. Dit artikel is niet van toepassing op het verveelvoudigen van een met elektronische middelen toegankelijke verzameling als bedoeld in artikel 10, derde lid.

Dit is van belang omdat ook gebruikers/consumenten van deepfakes rechten hebben, die een rol spelen bij de uiteindelijk inkadering van deepfakes. De privékopieexceptie houdt twee belangrijke punten in met betrekking tot deepfakes. Ten eerste kunnen personen beelden downloaden en hergebruiken voor het maken van deepfakes zonder aan het auteursrecht geboden te zijn, zo lang zij de deepfake slechts in de privésfeer gebruiken. Ten tweede kunnen burgers ook bestaande deepfakes downloaden, zolang zij die maar binnen de privésfeer bekijken of anderszins benutten en daar geen commerciële doeleinden mee nastreven. Wel is van belang dat er binnen het EU-recht nadere regels zijn gesteld aan de privékopieexceptie.

3.5 AI Regulering

Momenteel is er binnen de EU een voorstel voor een AI-regulering aanhangig. Omdat de definitieve versie nog een tijd op zich zal laten wachten en er nog de nodige wijzingen zullen plaatsgrijpen zal hier slechts kort de belangrijkste voorgestelde regels die op deepfakes van toepassing zijn worden besproken. Daarbij valt op dat de Commissie, in het voorlopige voorstel, kiest om deepfakes niet als zodanig te verbieden, noch geheel noch voor specifieke toepassing, maar slechts transparantievoorwaarden aan de publicatie en verspreiding van deepfakes te koppelen. Wel is het mogelijk dat bepaalde deepfake-apps of toepassingen onder de verboden of hoog-risico categorie van AI-systemen zullen vallen, maar duidelijk is dat op dit moment niet.[303]

Titel IV van de voorgestelde Regulering heeft betrekking op bepaalde AI-systemen met specifieke risico's voor manipulatie. Transparantieverplichtingen zullen gelden voor systemen die i) interactie hebben met mensen, ii) worden gebruikt om emoties te detecteren of associatie met (sociale) categorieën te

bepalen op basis van biometrische gegevens, of iii) inhoud genereren of manipuleren ("deep fakes"). Indien een AI-systeem wordt gebruikt om beeld-, audio- of video-inhoud te genereren of te manipuleren die merkbaar lijkt op authentieke inhoud, moet er een verplichting zijn om bekend te maken dat de inhoud op geautomatiseerde wijze is gegenereerd, behoudens uitzonderingen voor legitieme doeleinden (rechtshandhaving, vrijheid van meningsuiting). Dit stelt personen in staat geïnformeerde keuzes te maken of uit een bepaalde situatie terug te treden.

Overweging 70 van de Regulering stelt onder meer dat 'users, who use an AI system to generate or manipulate image, audio or video content that appreciably resembles existing persons, places or events and would falsely appear to a person to be authentic, should disclose that the content has been artificially created or manipulated by labelling the artificial intelligence output accordingly and disclosing its artificial origin.' Artikel 52 lid 3 stelt: 'Users of an AI system that generates or manipulates image, audio or video content that appreciably resembles existing persons, objects, places or other entities or events and would falsely appear to a person to be authentic or truthful ('deep fake'), shall disclose that the content has been artificially generated or manipulated. However, the first subparagraph shall not apply where the use is authorised by law to detect, prevent, investigate and prosecute criminal offences or it is necessary for the exercise of the right to freedom of expression and the right to freedom of the arts and sciences guaranteed in the Charter of Fundamental Rights of the EU, and subject to appropriate safeguards for the rights and freedoms of third parties.'

Dit betekent dat er wordt gekozen voor een transparantieverplichting en er, in het voorlopige voorstel van de Commissie, geen nadere regels of verplichtingen worden opgelegd aan deepfakes. Ook zijn er nog tal van vragen die moeten worden beantwoord gedurende het wetgevingstraject, zoals, maar niet beperkt tot: (1) De bepaling lijkt betrekking te hebben op alle gemanipuleerde inhoud; het probleem is echter dat, afhankelijk van de definitie van manipulatie, elke communicatietechnologie vervormt. Videoservices hebben ingebouwde tools die huidtinten egaliseren, audioservices filteren hoge tonen automatisch weg, enz. De schatting is dat over 5 jaar meer dan 90% van alle online content in een of andere vorm zal worden gemanipuleerd. Belangrijk is dat de AI-wet ook verwijst naar gemanipuleerde content over objecten, plaatsen en gebeurtenissen. Is deze bepaling ook van toepassing op een zon met een smiley-face? (2) De bepaling houdt alleen in dat de gebruiker moet melden dat de inhoud is gemanipuleerd: maar aan wie moet hij dergelijke informatie bekendmaken? Het algemene publiek; de afgebeelde persoon; het platform waarop het is geplaatst? (3) Het is de gebruiker van het AI-systeem die de verplichting heeft om informatie over de manipulatie te melden; hoe moet hij dit doen? Via metacontent of in de content zelf? Hoe groot of klein moet de uitleg zijn? Wat moet die informatie inhouden: 'deze inhoud is gemanipuleerd' of een beschrijving van wat er precies is gemanipuleerd en hoe? (4) Hoe deze informatie moet worden gecommuniceerd blijft onduidelijk. Het artikel spreekt slechts van ' shall disclose' terwijl de overweging spreekt van 'by labelling'. Labellen wordt vaak gebruikt binnen de AI context voor het prepareren van data door de data te labellen, dat wil zeggen te categoriseren. Dit zou

betekenen dat het gaat om metadata die wordt toegevoegd aan de eigenlijke data; hierdoor zou niet het algemene publiek op de hoogte worden gesteld van het feit dat content gemanipuleerd is (tenzij een automatische manipulatiedetectie programma op hun computer hebben draaien), maar kunnen wel platforms als Facebook en Google gemakkelijk content blokkeren die gemanipuleerd is. Dit heeft evenwel als mogelijk nadeel dat echte content ook gemanipuleerd kan lijken door er sporen van manipulatie aan toe te voeren. Een AI systeem 'leert' dan mogelijk ten onrechte dat een bepaalde video, audiofragment of andere content onecht is, terwijl dat niet het geval is. (5) Hoe verhoudt deze bepaling zich tot de AVG? Enerzijds benadrukt de AVG onder meer dat de verwerkte gegevens correct en actueel moeten zijn (principe van datakwaliteit); moet deze bepaling worden gezien als een verbod op deepfakes en zo ja, moet de AI-wet worden gezien als een uitzondering daarop? En wat voegt de AI-wet toe aan de transparantieverplichting als vervat in de AVG?

3.6 Onrechtmatige Daad

Het is van belang de onrechtmatigedaadsdoctrine te bespreken, omdat dit in horizontale relaties, dat wil zeggen bij burgers onderling, de belangrijkste grond zal zijn voor rechtszaken en procedures. Bij een schending van het civiel recht, de AVG of het strafrecht kan de burger een vordering instellen op basis van een onrechtmatige daad. Ook kan een dergelijke vordering als er bijvoorbeeld maatschappelijke normen zijn overschreden door de vervaardiging van een deepfake. In hoeverre kan een persoon gebruik maken van het onrechtmatige-daadsrecht ten aanzien van deepfakes? Aan welke voorwaarden moet worden voldaan om deze doctrine met succes te kunnen inroepen? Deze vragen worden toegelicht in deze paragraaf.

Hoofdstuk 11 van de Auteurswet is gewijd aan de handhaving van het auteursrecht. De Auteurswet bevat hierover een aantal bijzondere bepalingen en voor wat betreft de civielrechtelijke handhaving is het overige aangewezen op het civiele sanctie- en procesrecht, zoals dat is neergelegd in het BW en het wetboek van Burgerlijke Rechtsvordering. Een inbreuk op iemands portretrecht is naar Burgerlijk recht een onrechtmatige daad, hierdoor heeft de geportretteerde de mogelijkheid zich te beroepen op schending van het portretrecht indien hij een redelijk belang heeft om zich te verzetten tegen openbaarmaking. Door de ingebouwde belangenafweging in deze bepaling, kan er uiteindelijk sprake zijn van een schending van het portretrecht.

Naast het portretrecht kan er ook in algemene zin sprake zijn van een onrechtmatige daad.[304] Het onrechtmatige-daadsrecht fungeert als achtervang, indien een beroep op artikel 21 Auteurswet niet slaagt, staat een beroep op artikel 6:162 BW nog wel open.[305] Bij een beoordeling of een publicatie daadwerkelijk onrechtmatig is, zal altijd een belangenafweging plaatsvinden. Hierbij zullen alle omstandigheden van het geval een belangrijke rol vervullen. Het onrechtmatige-daadsrecht is ook van belang voor privacyschendingen in horizontale verhoudingen. Als een burger een deepfake van een andere burger maakt en daarmee ofwel de AVG ofwel Artikel 8 EVRM schendt, kan de andere burger schadevergoeding vragen middels een beroep op het BW. Hetzelfde geldt voor een strafrechtelijke veroordeling, die kan worden gebruikt om een veroordeelde persoon ook civielrechtelijk aan te spreken.

Voor aansprakelijkheid in de zin van artikel 6:162 BW zal aan een vijftal eisen moeten worden voldaan, te weten: onrechtmatige daad, toerekenbaarheid van de daad aan de dader, schade, causaal verband tussen de daad en de schade en als laatste relativiteit. Een van de eerste en tevens belangrijke arresten ten aanzien van de onrechtmatige daad, is HR Lindebaum/Cohen.[306] Dit arrest was het startpunt voor handelingen die in strijd met de zorgvuldigheid die in het maatschappelijk verkeer wordt betaamd, omdat sindsdien deze ook onder het begrip onrechtmatige daad vallen. Maatschappelijke onbetamelijkheid kan bij deepfakes een grote rol van betekenis spelen.

Onder onrechtmatige daad wordt begrepen: een inbreuk op een recht, een doen of nalaten in strijd met een wettelijke plicht en een doen of nalaten in strijd met hetgeen volgens ongeschreven recht in het maatschappelijk verkeer betaamt. Bovendien moet deze toerekenbaar aan de dader zijn, hiervan is sprake indien de dader schuld heeft aan de gedraging of wanneer een oorzaak voor zijn rekening komt. Hierbij gaat het enkel om het toerekenen van de daad, niet van de schade. Hiervoor is voldoende dat de dader schuld heeft aan zijn gedraging. Deze schuldvraag moet in dit verband worden opgevat als verwijtbaarheid. Om aansprakelijkheid te kunnen vestigen, moet tevens sprake zijn van enige geleden schade. Deze schade bestaat uit vermogensschade en ander nadeel, hier wordt immateriële schade mee bedoeld. Tussen de daad en schade moet daarnaast een causaal verband bestaan. De benadeelde heeft ten aanzien van deze causaliteit een stelplicht en moet dus zo nodig aannemelijk maken dat er sprake is van een causaal verband. Als laatste vereiste bestaat de relativiteit, die in artikel 6:163 BW wordt genoemd. Dit wil zeggen dat de norm die is overtreden door de dader moet zijn geschreven ter bescherming van het geschonden belang. Van belang hierbij is het Tandartsen-arrest[307] dat in 1958 door de Hoge Raad is gewezen, hiermee heeft de Hoge Raad voor het eerst de correctie-Langmeijer toegepast. Dit arrest ging over een tandarts, die zonder vergunning werkte. Het was niet toegestaan om zonder vergunning te werken, maar deze norm was niet geschreven om andere tandartsen te beschermen tegen deze vorm van concurrentie, maar de norm was ervoor bedoeld om patiënten te beschermen tegen een ondeskundige tandarts. Het relativiteitsbeginsel stond zodoende voor toewijzing van de vordering van de andere tandartsen in de weg.

Wanneer voldaan is aan deze vijf vereisten kan worden gesproken van aansprakelijkheid voor een onrechtmatige daad en kan zodoende schadevergoeding worden toegewezen. Een slachtoffer zou een schadevergoeding kunnen eisen op grond van artikel 6:162 BW, omdat er bijvoorbeeld sprake is van reputatieschade dan wel andere immateriële schade, zoals dat in artikel 6:106 BW wordt genoemd. Problematisch hieraan is wel dat het slachtoffer uit eigen initiatief een rechtszaak zal moeten starten, dat veel tijd, moeite en middelen kan kosten en dat internet providers niet altijd meewerken om de gegevens van plegers van onrechtmatige daden op te sporen en juridisch aan te kunnen spreken. Op dit punt zal meer uitgebreid worden stilgestaan in een volgend hoofdstuk (hoofdstuk 5). Indien een rechter tot de conclusie komt dat er sprake is van een onrechtmatige publicatie, dan kan daarop een verbod worden gevorderd en een rectificatie. Hierbij is het mogelijk om een dwangsom dan wel een schadevergoeding te vorderen.

3.7 Conclusie

Het 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. Verreweg de meeste financieel-economische en reputatie-gerelateerde uitingsdelicten zijn immers voldoende technologie-neutraal geformuleerd wat betreft de vorm waarin deze kunnen worden gepleegd. Wanneer deepfake-seksvideo's echter niet worden verspreid, maar puur voor eigen gebruik worden gemaakt en bekeken, valt dit niet onder een strafbepaling. Het is een rechtspolitieke vraag of het voor eigen gebruik maken van zulke deepfakes van iemand zonder diens toestemming strafbaar zou moeten worden gesteld. De enige mogelijke lacune in de wetgeving betreft het gat 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 als nadeel kan ontstaan. 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. Voor deepfakes levert dit niet per se een lacune in de rechtsbescherming op, aangezien zoals gezegd de meeste strafbepalingen door hun technologie-neutrale formulering van toepassing zijn. Mocht de wetgever het echter 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'. Hierdoor zou immers een algemene strafbaarstelling ontstaan van misbruik van iemands gelaat of stem als daaruit enig nadeel kan ontstaan.

Deepfakes lopen tegen een aantal obstakels op 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, wat ook 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. Wel moet worden bedacht dat er uitzonderingen kunnen bestaan in de vorm van de huishoudelijke exceptie en de verwerking van gegevens in het kader van de vrijheid van meningsuiting. Hoe nauw of wijd deze uitzonderingen dienen te worden geïnterpreteerd in de context van deepfakes is echter niet op voorhand en in het algemeen vast te stellen. Belangrijk is dat de AI-regulering, zoals voorgesteld door de Commissie, wel een specifieke bepaling bevat aangaande deepfakes, maar dat die slechts een transparantieverplichting betreft die voortbouwt op wat reeds volgt uit de AVG. Er zijn nog veel onduidelijkheden omtrent deze bepaling.

Binnen het Europees Verdrag voor de Rechten van de Mens moet er 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, goede naam en reputatie. Ook heeft het Hof bepaald 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, beoordeeld, kan niet in algemene zin worden gezegd hoe deze twee rechten zich bij deepfaketoepassingen tot elkaar verhouden. Dit zal per zaak moeten worden bekeken. Wel zijn twee bijzondere punten van belang. Enerzijds heeft het Hof een uitgebreide doctrine aangaande wat het noemt de ‘reasonable expectation of privacy’; daaruit volgt dat mensen ook in een werkomgeving of in de publieke ruimte mogen verwachten dat hun privacy wordt beschermd. Zelfs als een persoon extreme seksuele afbeeldingen van zichzelf op het internet zet, dan nog mag hij verwachten dat zijn privacy door anderen wordt gerespecteerd. Dat is van belang omdat hieruit volgt dat het recht op privacy op veruit het meeste materiaal dat wordt gebruik voor het genereren van deepfakes van toepassing zal zijn. Anderzijds geldt er een speciale doctrine voor bekende personen. Dit is van belang omdat de meeste deepfakes worden gemaakt ofwel over personen in de directe omgeving van de maker ofwel over bekende personen. Deze bekende personen moeten, zeker als het politici of ambtsdragers betreft, volgens het Hof meer dulden in termen van inperkingen in hun privésfeer en hun reputatie, eer en goede naam dan gewone burgers. Toch heeft het Hof eveneens benadrukt dat dergelijke inperkingen alsnog proportioneel dienen te zijn en dat ook publieke personen een recht op privacy toekomt.

Voetnoten hoofstuk 3

- 154 <https://www.politie.nl/themas/whatsapp-fraude-vriend-in-noodfraude.html>.
- 155 <https://www.fraudehelpdesk.nl/ondernemers-fraude/ceo-fraude/>.
- 156 Van der Velden & De Jonge, T&C Sr (Kluwer online), commentaar op art. 317, aant. 9.
- 157 Deze bepalingen stellen ook zogenoemde strafrechtelijke identiteitsfraude strafbaar.
- 158 *Kamerstukken II* 2011/12, 33 352, nr. 3, p. 24.
- 159 *Kamerstukken II* 2012/13, 33 352, nr. 7, p. 2.
- 160 *Kamerstukken II* 2011/12, 33 352, nr. 3.
- 161 *Kamerstukken II* 2012/13, 33 352, nr. 7, p. 2.
- 162 Ibid.
- 163 *Kamerstukken II* 2011/12, 33 352, nr. 3, p. 21.
- 164 Ibid.
- 165 HR 15 januari 1991, NJ 1991/668 (Rotterdamse computerfraude).
- 166 Zie o.a. EHRM, Pfeifer v. Austria App no 12556/03, 15 November 2007. EHRM, Rothe v. Austria App no 6490/07, 04 December 2012. EHRM, A. v. Norway App no 28070/06, 09 April 2009.
- 167 Bij zogenoemde culpoze delicten, waarvoor vereist is dat iemand aanmerkelijk nalatig is geweest, ontbreekt opzet. Wij zien geen culpoze delicten waarvoor deepfakes specifieke relevantie hebben, al valt natuurlijk niet uit te sluiten dat bijvoorbeeld een geval waarin een hartpatiënt een fatale hartaanval krijgt bij het zien van een afpersings-deepfakefilmpje waarin haar zoon wordt onthoofd, zou kunnen leiden tot vervolging voor dood door schuld (art. 307 Sr).
- 168 Ten Voorde, T&C Sr (Kluwer online), commentaar op art. 137e Sr, aant. 10 onder b en i.
- 169 Ten Voorde, T&C Sr (Kluwer online), commentaar op art. 137c Sr, aant. 9 onder c.
- 170 Janssens, T&C Sr (Kluwer online), commentaar op art. 270 Sr, aant. 1.
- 171 Vgl. HR 3 december 2013, ECLI:NL:HR:2013:1556, waarin een filmpje met seksuele gedragingen als smaadschrift wordt gekwalificeerd.
- 172 Janssens, T&C Sr (Kluwer online), commentaar op art. 261 Sr, aant. 9 onder h.
- 173 Vgl. HR 3 december 2013, ECLI:NL:HR:2013:1556.
- 174 Janssens, T&C Sr (Kluwer online), commentaar op art. 271 Sr, aant. 8.
- 175 Janssens, T&C Sr (Kluwer online), commentaar op art. 266 Sr, aant. 8 onder d.
- 176 HR 10 november 2015, ECLI:NL:HR:2015:3247.
- 177 *Kamerstukken II* 2018/19, 35 080, nr. 3, p. 22.
- 178 Ibid.
- 179 Ibid.
- 180 Ibid.
- 181 *Kamerstukken II* 2018/19, 35 080, nr. 3, p. 23.
- 182 Consultatieversie wetsvoorstel seksuele misdrijven (Wijziging van het Wetboek van Strafrecht en andere wetten in verband met de modernisering van de strafbaarstelling van verschillende vormen van seksueel grensoverschrijdend gedrag (Wet seksuele misdrijven)), 8 maart 2021, beschikbaar op <https://www.internetconsultatie.nl/wetsvoorstelseksuelemisdrijven>.
- 183 Kool & Lestrade, T&C Sr (Kluwer online), commentaar op art. 240, aant. 7 onder a.
- 184 Van Elst, T&C Sr (Kluwer online), commentaar op art. 350a, aant. 10 onder a.
- 185 *Kamerstukken II* 1989/90, 21 551, nr. 3, p. 23.
- 186 <https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2007/wp136_en.pdf>.
- 187 Ibid.
- 188 Ibid.
- 189 Ibid.
- 190 Ibid.
- 191 Overweging 26 AVG.
- 192 Overweging 27 AVG.
- 193 Overweging 14 AVG.
- 194 Albahar, M., & Almalki, J. (2019). Deepfakes: Threats and countermeasures systematic review. Journal of Theoretical and Applied Information Technology, 97(22), 3242-3250.
- 195 Artikel 2 lid 2 sub c AVG.

- 196 Overweging 18 AVG.
- 197 HvJ EU 6 november 2003, C-101/01, ECLI:EU:C:2003:596 (Bodil Lindqvist), r.o. 47.
- 198 'Article 29 Data Protection Working Party, Opinion 5/2009 on online social networking, 01189/09/EN WP 163, 12 June 2009, Brussels', Ec.europa.eu 9 april 2020. Citaat vertaald door projectteam op verzoek van WODC.
- 199 HvJ EU 11 december 2014, C-212/13, ECLI:EU:C:2014:2428 (František Ryneš v Úřad pro ochranu osobních údajů), r.o. 33.
- 200 Artikel 3 Richtlijn 95/46/EG van het Europees Parlement en de Raad van 24 oktober 1995 betreffende de bescherming van natuurlijke personen in verband met de verwerking van persoonsgegevens en betreffende het vrije verkeer van die gegevens. Publicatieblad Nr. L 281 van 23/11/1995 blz. 0031 - 0050
- 201 Overweging 12 Rbp.
- 202 Annex 2 Proposals for Amendments regarding exemption for personal or household activities. The situation under Directive 95/46/EC <https://ec.europa.eu/justice/article-29/documentation/other-document/files/2013/20130227_statement_dp_annex2_en.pdf>.
- 203 Annex 2 Proposals for Amendments regarding exemption for personal or household activities The situation under Directive 95/46/EC <https://ec.europa.eu/justice/article-29/documentation/other-document/files/2013/20130227_statement_dp_annex2_en.pdf>.
- 204 Artikel 3 AVG.
- 205 Artikel 4 sub 7 AVG.
- 206 Artikel 5 lid 1 sub a AVG.
- 207 Artikel 15 AVG.
- 208 Article 29 Working Party Guidelines on transparency under Regulation 2016/679 https://ec.europa.eu/newsroom/article29/document.cfm?action=display&doc_id=51025
- 209 Artikel 13 lid 1 en 2 AVG.
- 210 Artikel 13 lid 3 AVG.
- 211 Artikel 14 lid 3 AVG.
- 212 Artikel 4 sub 9 AVG.
- 213 Artikel 14 lid 5 sub b AVG.
- 214 Artikel 5 lid 1 sub b AVG.
- 215 <https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf>.
- 216 Ibid.
- 217 Artikel 16 AVG.
- 218 Artikel 12 lid 5 en 6 AVG.
- 219 Artikel 19 AVG.
- 220 Artikel 5 lid 1 sub d en lid 2.
- 221 <https://ec.europa.eu/newsroom/article29/document.cfm?action=display&doc_id=51030>.

222 Artikel 7 AVG. Voor kinderen geldt nog een speciale regeling. Artikel 8 AVG.

- 223 Artikel 4 sub 11 AVG.
- 224 Zie o.a. ECLI:NL:RBGEL:2018:2737, Rechtbank Gelderland, 31-05-2018, 21-06-2018, C/05/334012 / FA RK 18-668. ECLI:NL:RBOVE:2017:3924, Rechtbank Overijssel, 18-09-2017, C/08/171421 / FA RK 15-1077.
- 225 Artikel 6 lid 1 sub f AVG.
- 226 Overweging 47 AVG.
- 227 Artikel 4 sub 10 AVG.
- 228 Overwerging 50 AVG.
- 229 Artikel 6 lid 4 AVG.
- 230 Overweging 50 AVG.
- 231 Tweede Kamer, vergaderjaar 2017–2018, 34 851, nr. 3.
- 232 Dove, E., & Chen, J. (2020). What Does it Mean for a Data Subject to Make their Personal Data "Manifestly Public"? An Analysis of GDPR Article 9 (2)(e).
- 233 Artikel 18 AVG.
- 234 Artikel 21 lid 1 AVG.
- 235 Artikel 17 AVG.
- 236 Handvest van de grondrechten van de Europese Unie 2012/C 326/02.
- 237 Leidraad van de Raad voor de Journalistiek December 2019. Bij deze bepalingen wordt we aangegeven dat: Het afwijken van deze norm kan worden gerechtvaardigd wanneer er evident sprake is van een misstand én wanneer dit noodzakelijk is om de desbetreffende kwestie aan de orde te stellen.
- 238 Working Party 29, 'Opinion 4/2007 on the concept of

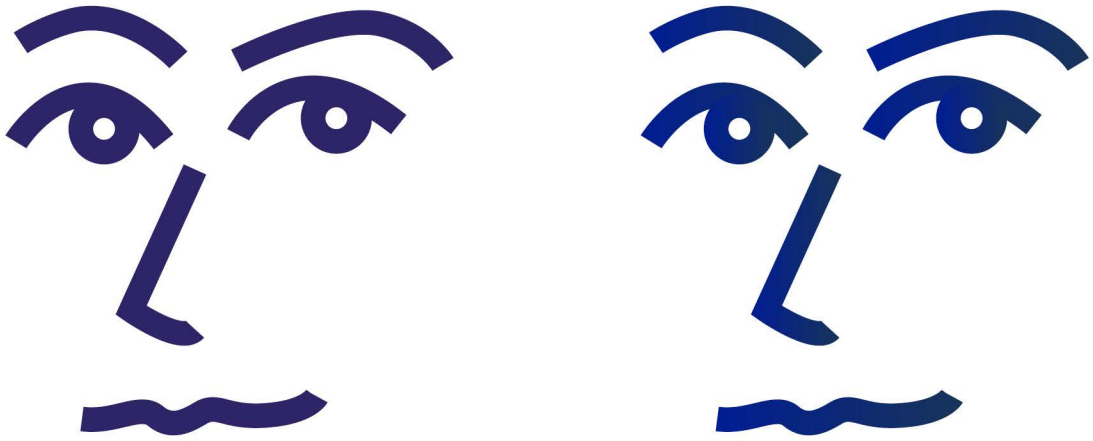
personal data', Brussels, 01248/07/EN, WP 136, 20 June 2007.

- 239 ECtHR, P.G. and J.H. v. UK, appl.no. 44787/98, 25 September 2001. ECtHR, Perry v. UK, appl.no. 63737/00, 17 July 2003. ECtHR, Klass a.o. v. Germany, appl.no. 5029/71, 6 September 1978. ECtHR, Malone v. UK, appl. no. 8691/79, 2 August 1984. ECtHR, Leander v. Sweden, appl.no. 9248/81, 26 March 1987. ECtHR, Köpke v. Germany, appl.no. 420/07, 5 October 2010. See in general: Mowbray, A. R. (2004). The development of positive obligations under the European Convention on Human Rights by the European Court of Human Rights' (Hart Publishing, Oxford). ECtHR, Copland v. UK, appl. no. 62617/00, 3 April 2007. ECtHR, Halford v. UK, appl. no. 20605/92, 25 June 1997. ECtHR, Peck v. UK, appl.no. 4467/98, 28 January 2003. ECtHR, Amann v. Switzerland, appl.no. 27798/95, 16 February 2000. ECtHR, Rotaru v. Romania, appl.no. 28341/95, 04 May 2000. ECtHR, Uzun v. Germany, appl.no. 35623/05, 2 September 2010. ECtHR, Hadzhiev v. Bulgaria, appl.no. 22373/04, 23 October 2012.
- 240 <https://www.echr.coe.int/Documents/FS_Data_ENG.pdf>.
- 241 ECtHR, Halford v. United Kingdom, application no. 20605/92, 02 March 1995. See also: ECtHR, Halford v. United Kingdom, application no. 20605/92, 18 April 1996.
- 242 Sychenko, E. & Chernyaeva, D. (2019) "The Impact of the ECHR on Employee's Privacy Protection" Italian Labour Law e-Journal 12.2.
- 243 ECtHR, Halford v. United Kingdom, application no. 20605/92, 25 June 1997.
- 244 ECtHR, P.G. and J.H. v. United Kingdom, application no. 44787/98, 25 September 2001.
- 245 ECtHR, Copland v. United Kingdom, application no. 62617/00, 03 April 2007.
- 246 ECtHR, Peev v. Bulgaria, application no. 64209/01, 26 July 2007. Zie ook het interessante vervolg hierop: ECtHR, Barbulescu v. Romania, application no. 61496/08, 12 January 2016. ECtHR, Barbulescu v. Romania, application no. 61496/08, 05 September 2017.
- 247 ECtHR, Steeg v. Germany, application nos. 9676/05, 10744/05 and 41349/06, 03 June 2008.
- 248 ECtHR, Antovic and Mirkovic v. Montenegro, application no. 70838/13, 28 November 2017.
- 249 ECtHR, Pay v. United Kingdom, application no. 32792/05, 16 September 2008. .
- 250 ECtHR, Benedik v. Slovenia, application no. 62357/14, 24 April 2018.
- 251 ECmHR, Asociacion de Aviadores de la Republica, Mata et al. v. Spain, appl.no. 10733/84, 11 March 1985. ECtHR, Saltuk v. Turkey, appl.no. 31135/96, 24 August 1999.
- 252 ECtHR, Marlow v. UK, appl.no. 42015/98, 5 December 2000.
- 253 ECmHR, Rayayd and Unanua v. Spain, appl.no. 31477/96, 15 January 1997. ECtHR, Minelli v. Switzerland, appl.no. 14991/02, 14 June 2005. ECtHR, GourGuénidzé v. Georgia, appl.no. 71678/01, 17 October 2006. ECtHR, Von Hannover v. Germany, appl.no. 59320/00, 24 June 2004. ECtHR, L. L. v. France, appl.no. 7508/02, 10 October 2006.
- 254 ECmHR, N. v. Sweden, appl.no. 11366/85, 16 November 1986.
- 255 ECtHR, Pfeifer v. Austria, appl.no. 12556/03, 15 November 2007.
- 256 ECtHR, Rothe v. Austria, appl.no. 6490/07, 04 December 2012.
- 257 ECtHR, A. v. Norway, appl.no. 28070/06, 09 April 2009.
- 258 ECtHR, Pretty v. the UK, appl.no. 2346/02, 29 April 2002.
- 259 ECtHR, Bogomolova v. Russia, appl.no. 13812/09, 20 June 2017.
- 260 ECtHR, Dink v. Turkey, appl.nos. 2668/07, 6102/08, 30079/08, 7072/09 and 7124/09, 14 September 2010.
- 261 ECtHR, Leroy v. France, appl.no. 36109/0302 October 2008.
- 262 ECtHR, Nikowitz and Verlagsgruppe News GMBH v. Austria, appl.no. 5266/03, 22 Frebruary 2007.
- 263 ECtHR, Handyside v. the UK, appl.no. 5493/72, 07 December 1976.

- 264 <https://www.echr.coe.int/Documents/Guide_Art_10_ENG.pdf>.
- 265 See further: Doherty, M. (2007). Politicians as a species of" Public Figure" and the Right to Privacy. Humanitas Journal of European Studies, 1(1), 35-56.
- 266 ECtHR, Von Hannover v. Germany, appl.no. 59320/00, 24 June 2004, § 69
- 267 ECtHR, Standard Verlag GMBH v. Austria (No. 2), appl.no. 21277/05, 04 June 2009. See also the dissenting opinion in: ECtHR, Verlagsgruppe News GMBH v. Austria (No. 2), appl.no. 10520/02, 14 December 2006.
- 268 ECtHR, Axel Springer AG v. Germany, application no. 39954/08, 07 February 2012. ECtHR, Von Hannover (2) v. Germany, appl.nos. 40660/08 and 60641/08, 07 February 2012. ECtHR, Rothe v. Austria, appl.no. 6490/07, 04 December 2012. ECtHR, Küchl v. Austria, appl.no. 51151/06, 04 December 2012. ECtHR, Verlagsgruppe Nes GMBH and Bobi v. Austria, appl.no. 59631/09, 04 December 2012. ECtHR, Węgrzynowski and Smolczewski v. Poland, appl.no. 33846/07, 16 July 2013. ECtHR, Ristamäki and Korvola v. Finland, appl.no. 66456/09, 29 October 2013. ECtHR, Lillo-Stenberg and Saether v. Norway, appl.no. 13258/09, 16 January 2014. ECtHR, Couderc and Hachette Filipacchi Associes v. France, appl.no. 40454/07, 10 November 2015. ECtHR, Ernst August von Hannover v. Germany, appl.no. 53649/09, 19 February 2015. ECtHR, Sousa Goucha v. Portugal, appl.no. 70434/12, 22 March 2016. ECtHR, Satakunnan Markkinapörssi Oy and Satamedia Oy v. Finalnd, appl.no. 931/13, 27 June 2017. ECtHR, Halldórsson v. Iceland, appl.no. 44322/13, 04 July 2017. ECtHR, Petkeviciute v. Lithuania, appl.no. 57676/11, 27 February 2018. ECtHR, M.L. and W.W. v. Germany, appl.nos. 60798/10 and 65599/10, 28 June 2018. ECtHR, Faludy-Kovács v. Hungary, appl.no. 20487/13, 23 January 2018.
- 269 ECtHR, Bohlen v. Germany, appl.no. 53495/09, 19 February 2015.
- 270 ECtHR, Ruusunen v. Finland, appl.no. 73579/10, 14 January 2014.
- 271 ECtHR, Ojala and Etukeno Oy v. Finland, appl.no. 69939/10, 14 January 2014. ECtHR, Sulamaki v. Finland, appl.no. 23605/09, 29 April 2014.
- 272 ECtHR, Alpha Doryforiki Teleorasi Anonymi Etairia v. Greece, appl.no. 72562/10, 22 February 2018.
- 273 ECtHR, Khadija Ismayilova v. Azerbaijan (no. 3), appl. no. 35283/14, 07 May 2020.
- 274 Van der Sloot, B. (2021). 'Expectations of Privacy'. In D. Hallinan et al. (eds), 'CPDP conference book', Hart.
- 275 ECtHR, Reklos and Davourlis v. Greece, appl.no. 1234/05, 15 januari 2009.
- 276 GS Onrechtmatige daad VII.8.2.6
- 277 HR 21 januari 1994, ECLI:NL:HR:1994:ZC1240 (Ferdi E.)
- 278 HR 21 januari 1994, ECLI:NL:HR:1994:ZC1240 (Ferdi E.), r.o. 3.5.
- 279 T& C IE, commentaar op art. 21 Aw
- 280 HR 30 oktober 1987, ECLI:NL:HR:1987:AD0034, NJ 1988/277
- 281 HR 2 mei 2003, ECLI:NL:HR:2003:AF3416, NJ 2004/80
- 282 HR 14 juni 2013, ECLI:NL:HR:2013:CA2788 (Cruijf/Tirion), r.o. 3.4.
- 283 HR 1 juli 1988, ECLI:NL:HR:1988:AB7688.
- 284 HR 19 januari 1979, ECLI:NL:HR:1979:AC6461, NJ 1979/383.
- 285 HR 14 juni 2013, ECLI:NL:HR:2013:CA2788, r.o. 3.6.2.
- 286 HR 14 juni 2013, ECLI:NL:HR:2013:CA2788 m.nt. P.B. Hugenholtz.
- 287 GS Onrechtmatige daad VII.8.3.3.5
- 288 Rb. Breda 24 juni 2005, AMI 2005/5 nr. 14 (Katja Schuurman).
- 289 Rb. Breda 24 juni 2005, AMI 2005/5 nr. 14 (Katja Schuurman), r.o. 3.19.
- 290 Rb. Breda 24 juni 2005, AMI 2005/5 nr. 14 (Katja Schuurman), r.o. 3.22.
- 291 Hof Amsterdam 2 juni 2020, ECLI:NL:GHAMS:2020:1410.
- 292 IER 2020/46, par. 7.1.
- 293 HR 30 oktober 1987, ECLI:NL:HR:1987:AD0034, NJ 1988/277

- 294 <http://www.beeldbalie.nl/portretrecht-en-kunst-werken>.
- 295 Rb. Amsterdam 2 mei 2018, ECLI:NL:RBAMS:2018:2983.
- 296 *Aanhangsel Handelingen II* 2014/15, 933,
- 297 *Aanhangsel Handelingen II* 2014/15, 933, p. 3.
- 298 HR 30 oktober 1987, ECLI:NL:HR:1987:AD0034, NJ 1988/277, r.o. 3.4.
- 299 HR 2 mei 1997, ECLI:NL:HR:1997:ZC2364 (Discodanser), r.o. 3.3.
- 300 Rb. Rotterdam 1 mei 2020, ECLI:NL:RBROT:2020:4296, r.o. 4.10.
- 301 Rb. Amsterdam 18 januari 2012, HA ZA 11-1621, r.o. 4.4 (B/ TomTom & M).
- 302 Rb. Midden-Nederland, 9 januari 2020, ECLI:NL:RBMNE:2020:24.
- 303 Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence and Amending Certain Union Legislative Acts {SEC(2021) 167 final} - {SWD(2021) 84 final} - {SWD(2021) 85 final}.
- 304 *Aanhangsel Handelingen II* 2017/18, nr. 828.
- 305 GS Onrechtmatige daad, VII 8.2.6
- 306 HR 31 januari 1919, ECLI:NL:HR:1919:AG1776.
- 307 HR 17 januari 1958, ECLI:NL:HR:1958:AG2051.

.



Nep of echt?

Nep of nepper?

Hoe onderscheiden we ‘nep’ van ‘echt’? En wat is de definitie van zowel ‘nep’ als ‘echt’ wanneer alle digitale content deels of volledig gemanipuleerd of gecreëerd is door (deepfake)technologie? Hoe herkennen we een manipulatie? En wat is het verschil tussen nep en nepper?

4. Procesrecht en procedurele vragen

In dit hoofdstuk wordt ingezoomd op één specifieke dreiging van deepfakes, omdat die raakt aan het functioneren van de rechtsstaat als zodanig, namelijk de introductie van deepfake-bewijsmateriaal in de rechtszaal. Eerst zal een korte schets van de problematiek worden gegeven (paragraaf 4.1). Dan volgt een analyse van de procesrechtelijke pijlers binnen het civiel recht (paragraaf 4.2) en het strafrecht (paragraaf 4.3). Het bestuursprocesrecht blijft hier achterwege omdat dat van marginaal belang zal zijn in conflicten tussen burgers onderling. Tot slot zal een korte conclusie worden gegeven (paragraaf 4.4).

4.1 Deepfakes in de rechtszaal

Of normering nu geschiedt via het strafrecht, het gegevensbeschermingsrecht, het civiel recht of middels zachtere normen en of handhaving nu geschiedt door burgers zelf, door internet intermediairs of van staatswege, altijd zal er een bewijsvraagstuk zijn. Dit bewijsvraagstuk gaat verder dan traditionele vragen in het recht, zoals naar de rechtmatigheid van een bepaald deepfake bericht. Het gaat om de bewijslast ten aanzien van de vraag of het deepfake betreft, dat wil zeggen of beelden of audiofragmenten authentiek zijn en zo ja, in welke mate.

Het zal soms moeilijk zijn om (voldoende overtuigend) te bewijzen dat een bepaalde (beweerde) deepfake-uiting – bijvoorbeeld een video waarin iemand racistische uitspraken lijkt te doen, wat mogelijk als smaad kan worden gekwalificeerd – daadwerkelijk een deepfake is. De bewijsbaarheid is van invloed op de vraag of nieuwe strafbaarstellingen wenselijk kunnen zijn: voorzienbare bewijsproblemen bij bestaande bepalingen kunnen immers een reden zijn om nieuwe strafbaarstellingen in te voeren, en voorzienbare bewijsproblemen bij mogelijk nieuwe bepalingen kunnen een argument zijn om af te zien van strafbaarstelling of om de formulering van een beoogde bepaling aan te passen.

Pfefferkorn onderscheidt bijvoorbeeld een aantal mogelijke bewijsrechtelijke problemen:[308]

- Met een deepfakevideo is er geen persoon die aanwezig was bij het moment dat de opname zou zijn genomen, omdat het incident niet is gebeurd;
- Als de video wordt aangedragen als bewijs door een van de partijen, wie moet dan bewijzen dat het materiaal authentiek is en welke standaard geldt daarvoor;
- Als een video online wordt geplaatst, dezelfde vraag: is het de plaatser van het bericht, degene die de authenticiteit betwist of het platform;
- Wat als materiaal niet direct wordt vrijgegeven, maar een video wordt 'gearchiveerd' en pas na iemands dood wordt vrijgegeven (bijvoorbeeld een compromitterende video om diens morele nalatenschap te ruïneren); wie kan dan de authenticiteit van een video betwisten en op welke grond?
- Welke gevolgen heeft deepfake voor waarlijk authentiek materiaal; mogelijk kan een verdachte/burger nu altijd claimen dat een video, beeld of audiofragment nep is, omdat het mogelijk is dat deze gefabriceerd is;
- Idem voor dat geval dat een internetprovider zich geconfronteerd ziet met twee burgers (uploader en de afgebeelde) met conflicterende claims ten aanzien van authenticiteit.

4.2 Civiel procesrecht

4.2.1 Algemeen

In een civiele procedure regelt het bewijsrecht het verzamelen en het beoordelen van bewijs. Bewijs is in feite alle informatie die relevant is voor de rechterlijke procedure. Door middel van het verzamelen van bewijs kan een rechter een zo volledig en correct mogelijke beslissing nemen. De rechter heeft als enige de taak het geleverde bewijs te beoordelen en of degene die de bewijslast heeft, in de waarheidsvinding is geslaagd. De basis voor wie de bewijslast heeft, ligt in hoofdregel 'wie eist, bewijst', gebaseerd op de zogenaamde objectiefrechtelijke leer. Deze hoofdregel houdt in dat degene die zich beroept op rechtsgevolgen van gestelde feiten of rechten die feiten of rechten ook moet bewijzen. Hieraan is een bewijsrisico verbonden, wat inhoudt dat de partij die de bewijslast heeft, het risico draagt dat de gestelde feiten niet bewezen worden.[309] Dit geldt ook in 'non-liquet-situaties', wat betekent dat de feiten niet bewezen kunnen worden, maar evenmin kan het tegendeel worden bewezen. Overigens ziet de bewijslast zowel op de bewijsvoeringslast als het bewijsrisico, in de praktijk vallen deze vaak samen.[310]

Voorafgaand aan de bewijslast, vindt de stelplicht plaats. Dit vindt zijn grondslag in artikel 150 Rv. Dit artikel houdt kort gezegd in dat degene die zich op een rechtsgevolg beroept een stelplicht heeft ten aanzien van de feiten die tot dat rechtsgevolg leiden; wie stelt zal ook moeten bewijzen. Die partij zal aan de rechter duidelijk moeten maken dat en waarom de rechter die feiten als vaststaand moet aannemen en aan zijn beslissing ten grondslag moet leggen.[311] Als er onvoldoende wordt gesteld, zal niet aan de bewijslevering worden toegekomen. Indien een partij meer heeft gesteld dan nodig is voor het gewenste rechtsgevolg, komt de bewijslast van het meerdere niet bij die partij te liggen.[312] Artikel 150 Rv zegt verder niets over wat welke partij moet stellen, het enige dat artikel 150 Rv eigenlijk bepaalt is dat degene die zich beroept op bepaalde rechtsgevolgen, daar ook de bewijslast voor draagt. De Hoge Raad heeft in zijn arrest van 25 september 2020[313] nogmaals op een rijtje gezet welke mogelijkheden de rechter heeft in het algemeen ten aanzien van de stelplicht en de bewijslast tussen partijen:

> *"1. Hoofdregel is dat de stelplicht en de bewijslast ter zake van het condicio sine qua non-verband op de benadeelde rusten (art. 150 Rv). Dit geldt ook bij besluitaansprakelijkheid.*
>
> *2. De rechter kan in de bijzondere omstandigheden van het geval aanleiding vinden om te oordelen dat op de wederpartij van de benadeelde een verzwaarde motiveringsplicht rust (dat wil zeggen een verplichting om voldoende feitelijke gegevens te verstrekken ter motivering van haar betwisting, teneinde de benadeelde voldoende aanknopingspunten te bieden voor het nader onderbouwen en zo nodig bewijzen van de door hem gestelde feiten).*
>
> *3. Als een partij haar betwisting van de stellingen van de andere partij onvoldoende motiveert, kan de rechter aan die betwisting voorbijgaan, zodat de gestelde feiten vaststaan.*
>
> *4. Daarnaast kan de rechter, indien de wederpartij de stellingen van de benadeelde ter zake van het condicio sine qua non-verband voldoende heeft betwist, op grond van zijn waardering van de wederzijdse stellingen en het voorhanden bewijsmateriaal de betwiste stelling voorshands bewezen achten, behoudens tegenbewijs.*

5. Ten slotte kan de rechter oordelen dat in de bijzondere omstandigheden van het geval uit de eisen van redelijkheid en billijkheid een omkering van de bewijslast voortvloeit (art. 150 slot Rv).”[314]

Hoewel hierboven ook de hoofdregel als uitgangspunt wordt genomen, waarbij de partij die stelt, de bewijslast ervan draagt, draagt de verweerder de bewijslast van al die feiten met rechtsgevolgen die de toewijzing van de vordering van de eiser blokkeren.[315] Die partij zal dus de feiten moeten bewijzen om het rechtsgevolg dat wordt ingeroepen ook daadwerkelijk te laten intreden, en die partij draagt ten aanzien van die stelling dan ook het bewijsrisico. Dit wordt ook wel een ‘zelfstandig’ of een ‘bevrijdend’ verweer genoemd, in meer simplistische taal wordt het ook wel aangeduid als een ‘ja, maar-verweer’. Met zo’n verweer erkent de verweerder wel dat de gestelde feiten juist zijn, maar beroept zich op een bevrijdende omstandigheid. Dit is bijvoorbeeld het geval geweest in een zaak waarbij het gaat om een geschil tussen een advocaat en diens cliënt. De cliënt stelt namelijk dat de advocaat tekort is geschoten en dus een beroepsfout heeft gemaakt, waardoor de cliënt een schadevergoeding eist. De advocaat erkent deze fout, maar stelt dat de cliënt even grote schade zou hebben geleden als de fout niet zou zijn gemaakt.[316]

4.2.2 Tegemoetkomingen ten aanzien van de bewijslast

In sommige gevallen kan een rechter de partij waarop de bewijslast rust tegemoetkomen. Dit kan in gevallen waarvan de rechter vermoed dat de stelling waar is, bijvoorbeeld op basis van de overgelegde stukken, maar dit kan ook op basis van een wettelijk of jurisprudentieel vermoeden. De stelling wordt dan voorshands bewezen. Hiertegen kan de wederpartij tegenbewijs leveren.[317] Hierbij is voldoende dat het bewijsvermoeden enkel wordt ontzenuwd, het tegendeel hoeft dus niet te worden bewezen. Wel zal deze partij – die het tegenbewijs wil leveren – voldoende moeten hebben gesteld om te worden toegelaten.[318]

Een andere bewijslastverdeling kan voortvloeien uit de eisen van redelijkheid en billijkheid of uit enig bijzondere regel. Deze andere bewijslastverdeling geschiedt alleen onder bijzondere omstandigheden en dient goed te worden gemotiveerd.[319] Dit is enkel mogelijk onder bijzondere omstandigheden, en indien toepassing van de hoofdregel leidt tot onbillijke resultaten en geen bijzondere – geschreven en ongeschreven – regel uitkomst biedt. Echter, de rechter is terughoudend in het gebruik maken van deze mogelijkheid.[320]

Als laatste tegemoetkoming, is er de zogenaamde ‘omkeringsregel’. Dit is een ingrijpende maatregel en hierbij zal de partij die een bepaald rechtsgevolg inroept niet hoeven te bewijzen, maar de wederpartij zal moeten bewijzen dat die feiten zich niet hebben voorgedaan. Deze ziet toe op de bewijslast van het causale verband (condicio sine qua non) bij bijvoorbeeld letselschade en op een toerekenbare tekortkoming in de nakoming van een verbintenis. Met het oog op het onderwerp van dit onderzoek, zal – omdat deze tegemoetkoming geen toepassing zal hebben op dit onderwerp – worden volstaan met het enkel benoemen van deze tegemoetkoming.

4.2.3 Beginselen

Het bewijsrecht is omgeven van een aantal algemene beginselen, zoals het beginsel van hoor en wederhoor, neergelegd in artikel 19 Rv. Partijen in civiele procedures hebben naast rechten,

ook een aantal verplichtingen. Partijen zullen bijvoorbeeld feiten moeten aanvoeren die hun stellingen ondersteunen. Artikel 20 Rv waakt voor een onnodige vertraging in het proces, dit hangt samen met artikel 6 van het EVRM.[321] De precieze termijnen zijn vooralsnog onduidelijk, waarbij in sommige gevallen de waarheidsvinding prevaleert boven dat van een voortvarende rechtspleging.[322] Het bewijsrecht is onlosmakelijk verbonden met de waarheids- en volledigheidsplicht, neergelegd in artikel 21 Rv.[323] Partijen zijn verplicht alle relevante informatie aan te dragen. Op grond van artikel 21 Rv zijn de partijen dus verplicht de feiten, die voor de beslissing van belang zijn, volledig en naar waarheid aan te voeren. Dit kan met andere woorden de waarheidsplicht worden genoemd. Hierbij is goed om op te merken dat feiten die van belang kunnen zijn voor de beslissing, niet bewust mogen worden achtergehouden of onjuist mogen worden weergegeven. Gevolg hiervan kan zijn dat partijen feiten zullen moeten aandragen die niet in het belang zijn van henzelf, maar wel kunnen bijdragen aan het standpunt van de wederpartij. Dit kan in gevallen waarbij de ene partij afhankelijk is van de andere partij, erg gunstig zijn. Ook staat in artikel 21 Rv dat als deze verplichting niet wordt nageleefd, de rechter daaruit gevolgtrekking kan maken die zij geraden acht. Dit houdt bijvoorbeeld in dat de rechter zich extra kritisch opstelt tegenover de beweringen van de partij. Dit kan mogelijk gevolgen meebrengen voor de bewijswaardering.

De rechter kan ook de bewijslast verzwaren. De rechter heeft de mogelijkheid om op grond van artikel 22 Rv partijen te bevelen een bepaalde stelling toe te lichten of bepaalde, op de zaak betrekking hebbende bescheiden over te leggen. Artikel 22a Rv bepaalt dat in geval van bewijsstukken die mogelijk de lichamelijke dan wel geestelijke gezondheid van een partij zal schaden, deze stukken enkel door de advocaat of arts worden bekeken. Hierbij is van belang dat het de gezondheid onevenredig zou schaden, het kan hierbij ook gaan om bedrijfsgeheimen. Op grond van artikel 22b Rv dienen partijen te specificeren welke stelling de door haar in het geding gebrachte stukken zijn bedoeld en welk onderdeel daarvan van belang is voor het proces.[324] Als de indiener dit niet duidelijk maakt, kan de rechter de overgelegde stukken buiten beschouwing laten. Verder bepaalt artikel 24 Rv dat de rechter de zaak onderzoekt en beslist op basis van hetgeen partijen aan hun vordering, verzoek of verweer ten grondslag hebben gelegd. Dit heeft een positieve en negatieve zijde. De negatieve zijde houdt in dat de rechter feitelijke gronden van de vordering, het verzoek of het verweer niet mag aanvullen.[325] De positieve zijde houdt in dat de rechter moet onderzoeken of de feiten die in de procedure zijn aangevoerd en komen vast te staan, het gevorderde ook kunnen dragen. Artikel 24 Rv biedt een juridische basis aan de stelplicht van partijen.

4.2.4 Betrouwbaarheid van bewijs

Het uitgangspunt bij bewijs is dat het kan worden geleverd door alle middelen (artikel 152 Rv). Zodoende kunnen alle informatiebronnen gebruikt worden als bronnen van informatie op basis waarvan het waarheidsgehalte van andere bronnen kan worden vastgesteld. Bewijs dient namelijk zowel feitelijke beweringen, maar ook de betrouwbaarheid, authenticiteit of integriteit van informatie. Middels bewijs kan ander bewijs bijvoorbeeld ontzenuwd worden.[326]

Authenticiteit van een bewijsmiddel houdt eigenlijk in of de herkomst van het bewijsmiddel of de informatie vaststaat. Onder integriteit

wordt verstaan dat het bewijsmiddel als zodanig en de inhoud onaangetast zijn door onbevoegde manipulatie.[327]

Het recht kent zelf ook maatregelen om de betrouwbaarheid van een bewijsmiddel te verhogen, in geval van een getuige zal de eed of de belofte de integriteit van de informatie meer waarborgen, maar ook de ondertekening van geschriften dient de authenticiteit van het bewijsmiddel: het is opgemaakt door degenen die het hebben ondertekend. Ten aanzien van een elektronische handtekening heeft bijvoorbeeld artikel 3:15a BW bepaalde eisen opgesteld die de integriteit en de authenticiteit net zo waarborgen als de handgeschreven handtekening. Dit alles waarborgt wellicht niet honderd procent de authenticiteit en integriteit van het bewijsmiddel, maar als uitgangspunt geldt dat dit niet voor het bewijs doorslaggevend is. Partijen zullen actief de authenticiteit of integriteit ter discussie moeten stellen, anders zal de rechter geen aanleiding zien om het bewijsmiddel ter onderzoek te stellen. Dit geldt voor alle categorieën van bewijsmiddelen.[328]

Een rechter zal en hoeft dus niet preventief onderzoek te doen naar de betrouwbaarheid van een bewijsmiddel. Dit zal hij alleen hoeven doen als een van de partijen de authenticiteit of integriteit van het bewijsmiddel ter discussie stelt. Ten aanzien van elektronische gegevens wordt in de literatuur gesteld dat in beginsel de bestaande eisen voor de authenticiteit en integriteit afdoende moeten zijn, maar dat in verband met de moeilijkheden die kunnen ontstaan door elektronische gegevens, de rechtspraak of de wet nadere regels kunnen gaan stellen.[329]

4.2.5 Keuze voor onderzoek

Het bewijsrecht richt zich met name op de klassieke informatiebronnen, zoals het geschrift, de verklaringen van getuigen en de deskundige. Daar zijn wel de elektronische gegevens bij gekomen. Om een volledig beeld te krijgen van de regels rondom bewijs, zullen eerst de drie klassieke categorieën worden besproken, daarna de elektronische gegevens, om vervolgens de opgedane kennis toe te passen op deepfakes.

4.2.6 Schriftelijke akte

Er is altijd geprobeerd middelen te vinden om de authenticiteit en integriteit van bijvoorbeeld schriftelijke bewijsmiddelen te verifiëren.[330] De betrouwbaarheid moet volgens het bestaande bewijsrecht, zoals dat is neergelegd in artikelen 149-200 Rv, worden behandeld. Meestal zal – in geval van een geschrift – het bewijsmiddel worden onderzocht door een of meer deskundigen. Indien niet met voldoende zekerheid gesteld kan worden dat er inderdaad sprake is van een vervalst bewijsstuk, zal de rechter moeten beslissen voor rekening van welke partij dat komt. Hiervoor wordt gekeken in artikel 159 Rv, dat twee bepalingen kent. In lid 1 is bepaald dat als de uiterlijke schijn meebrengt dat op voorhand wordt gedacht dat het bewijsstuk authentiek is, als zodanig geldt behoudens bewijs dat het tegendeel bewijst. Dit komt overeen met de hoofdregel van artikel 150 Rv. Lid 2 richt zich op de onderhandse akte. In het geval van een onderhandse akte rust de bewijslast op degene die de akte als bewijsstuk gebruikt of zich daarop beroept. Bovendien kan de rechter, als de valsheid van zo'n akte blijkt uit bijvoorbeeld onregelmatigheden, een vermoeden van valsheid putten, maar als dat niet aan de orde komt, zal degene die de valsheid stelt, het bewijsrisico ervan dragen en deze valsheid dus moeten bewijzen.[331]

4.2.7 Getuigenverklaring

Bij bewijs in de vorm van een getuigenverklaring speelt het geheugen – en de interpretatie ervan – een belangrijke rol. Bij een getuigenverklaring is het van belang in ogenschouw te nemen dat de waarneming niet de exacte werkelijkheid weergeeft. Een rechter kan niet weten of een verklaring juist is en zal dus de betrouwbaarheid en geloofwaardigheid zelf moeten inschatten. De betrouwbaarheid van een getuige is lastig in te schatten, er is geen empirische basis voor de veronderstelling dat bepaalde mensen niet waarheidsgetrouw zouden kunnen verklaren en daardoor geen betrouwbare getuige zijn. Hierdoor moet gekeken worden naar de geloofwaardigheid van de getuigenverklaring. Dit moet worden los gezien van de persoon zelf, omdat een persoon over het ene onderwerp geloofwaardig kan verklaren, terwijl dat over een ander onderwerp niet zo is. Er moet rekening gehouden worden met de mogelijkheid dat de verklaring berust op een onjuiste waarneming of een onjuiste herinnering. Herinneringen zijn namelijk aan veranderingen onderhevig. Getuigen die verklaringen op basis van onjuiste waarnemingen of onjuiste herinneringen zijn gevallen van 'dwalende getuigen'. De rechter zal de getuigenverklaringen van beide partijen tegen elkaar af moeten wegen. Bovendien moet de rechter zelf de getuigen horen, dit in tegenstelling tot het strafrecht. De rechter mag dus niet afgaan op schriftelijke verklaringen die in het dossier bevinden. Dit geeft de rechter ook de mogelijkheid vragen te stellen aan de getuigen, om zo een duidelijk beeld van de waarheid te krijgen. Verder is de rechter verplicht om op grond van artikel 177 lid 1 Rv in het proces-verbaal de verhouding van de getuige tot een van de partijen te noteren, zoals of er sprake is van een dienstverband dan wel een relatie van bloed- of aanverwantschap. Het weten van de verhouding tussen de getuige en de partijen, kan een inkijk geven in de belangen die de getuigen zelf kan hebben om een bepaalde verklaring te geven, maar uiteindelijk is dit onvoldoende om vast te kunnen stellen of een verklaring betrouwbaar is. Om de betrouwbaarheid vast te stellen is het perspectief van de rechter op de waarheidsgetrouwheid belangrijk. De Bock spreekt over een aantal criteria hiervoor, deze zullen hieronder kort worden besproken:[332]

- Relevantie van de verklaring: Belangrijk hierbij is of hetgeen verklaard is ook relevant is voor de bewijsopdracht. Het moet dus relevant zijn voor het te bewijzen feit. De rechter zal hierbij precies aan de getuige moeten vragen naar de wijze waarop hij betrokken is en waarop hij zijn kennis baseert;
- Consistentie van de verklaring: tegenstrijdigheden in verklaringen, doet afbreuk aan de waarheidsgetrouwheid van de verklaring. Een innerlijke tegenstrijdige verklaring zal geen positieve bijdrage leveren aan de inschatting van de waarheid door de rechter. Het tegenovergestelde is bovendien ook mogelijk, de overeenstemmingen in verklaringen van meerdere getuigen kan de schijn wekken van waarheidsgetrouwheid. Bovendien kan de consistentie in het verband met andere bewijsstukken worden gezien, dus komt de verklaring overeen met bijvoorbeeld brieven of e-mails die zijn verstuurd. Ook bij de vraag of er sprake is van consistentie, zal een rechter door moeten vragen en de getuige confronteren met geconstateerde inconsistenties.
- Kwaliteit van de verklaring: de mate van kwaliteit kan worden bepaald aan de hand van de kwaliteit van de waarneming van de getuige. Hierbij moet worden gedacht aan het gewicht of de betekenis van een getuigenverklaring voor

de bewijsopdracht. Heeft de getuige de situatie zelf waargenomen, of heeft hij het van horen zeggen? Overigens kunnen verklaringen 'van horen zeggen' wel worden meegenomen in het bewijs. De eigen waarneming is belangrijk, een gekregen indruk kan hier onder vallen. Ook in dit geval is het belangrijk dat de rechter door blijft vragen, waarbij ook weer rekening wordt gehouden met de positie van de getuige.

- Coherentie van de verklaring: dit gaat over de begrijpelijkheid. Een verklaring kan incoherent zijn als het bijvoorbeeld leemtes of onduidelijkheden bevat. Hierbij is de narratieve structuur van een verklaring essentieel, het moet een logisch verhaal bevatten. Wel dient de rechter rekening te houden met het feit dat niet iedereen in staat is een goed verhaal te vertellen, het kan bijvoorbeeld zijn dat een getuige de context waarover hij verklaart niet heeft begrepen. De rechter kan zich ook anticiperend opstellen door middel van het stellen van bepaalde (korte) vragen, waar de getuige wel een antwoord op kan geven.

Op basis van deze criteria komt een eenduidig beeld naar voren, belangrijk is – in ieder geval bij getuigenverklaringen – dat de rechter door blijft vragen en geen genoegen neemt met de enkele (eerste) verklaring van de getuige. Door door te vragen is het meer mogelijk om de betrouwbaarheid van een verklaring vast te stellen.

De Bock noemt nog een vijfde criterium, waar mogelijk ruimte voor vrijgemaakt moet worden: geloofwaardigheid van de verklaring. Hierbij kan onder meer gedacht worden aan de fysieke houding, de manier waarop vragen worden beantwoord, de lichaamstaal (zweten, blozen, knipperen met de ogen, trillende handen). Deze opvallendheden kunnen duiden dat de verklaring niet waarheidsgetrouw is, maar dit biedt geen zekerheid. Daardoor is het maar de vraag of deze factoren als indicatie kunnen dienen.

4.2.8 Deskundigenverklaring

De deskundige is op één lijn te stellen als de getuige. Hij is als ware een gespecialiseerde getuige. Uit onderzoek door de Raad voor de rechtspraak blijkt dat een rechter de betrouwbaarheid van het werk van een deskundige maar ten dele kan onderzoeken en dat de rechter en een deskundige te weinig over elkaars vakgebied weten om de betrouwbaarheid en validiteit van elkaars werk te kunnen beoordelen.[333] De rechtspraak kent maar beperkte motiveringseisen ten aanzien van het oordeel van de deskundige.[334] In geval van zaken over letselschade – maar dit uitgangspunt wordt ook buiten letselschadezaken gebruikt – waarbij partijen tezamen hebben besloten een deskundige in te schakelen om vragen te beantwoorden, zal de uitkomst van de deskundige het uitgangspunt zijn voor verdere afwikkeling. Om af te wijken van dit uitgangspunt, zullen zwaarwegende bezwaren tegen de rapportage moeten bestaan, wil een rechter het rapport buiten beschouwing houden. Bovendien neemt een rechter in de meeste gevallen het oordeel van de deskundige over, wat hij vervolgens summier motiveert.

4.2.9 Motiveringseisen

Een rechter zal zijn beslissingen zodanig moeten motiveren, dat zij voldoende inzicht geeft in de aan haar ten grondslag liggende gedachtegang om de beslissing.[335] In het geval van een deskundigenbericht, heeft de Hoge Raad[336] de motiveringsplicht als volgt beschreven:

> *"Vooropgesteld moet worden dat voor de rechter een beperkte motiveringsplicht geldt ten*

aanzien van zijn beslissing om de bevindingen van deskundigen al dan niet te volgen. Wel dient hij bij de beantwoording van de vraag of hij de conclusies waartoe een deskundige in zijn rapport is gekomen in zijn beslissing zal volgen, alle terzake door partijen aangevoerde feiten en omstandigheden in aanmerking te nemen en op basis van die aangevoerde stellingen in volle omvang te toetsen of aanleiding bestaat van de in het rapport geformuleerde conclusies af te wijken. Ingeval partijen, door zich te beroepen op de uiteenlopende zienswijzen van de door haar geraadpleegde deskundigen, voldoende gemotiveerde standpunten hebben ingenomen en voldoende duidelijk hebben aangegeven waarom zij het oordeel van een door de rechter benoemde deskundige al dan niet aanvaardbaar achten, geldt het volgende. Indien de rechter in een geval waarin de opinie van andere, door een der partijen geraadpleegde, deskundigen op gespannen voet staat met die van de door de rechter benoemde deskundige, de zienswijze van deze deskundige volgt, zal de rechter zijn beslissing in het algemeen niet verder behoeven te motiveren dan door aan te geven dat de door deze deskundige gebezigde motivering hem overtuigend voorkomt. Wel zal de rechter op specifieke bezwaren van partijen tegen de zienswijze van de door hem aangewezen deskundige moeten ingaan, als deze bezwaren een voldoende gemotiveerde betwisting inhouden van de juistheid van deze zienswijze. Volgt de rechter echter de zienswijze van de door hem benoemde deskundige niet, dan gelden in beginsel de gewone motiveringseisen en dient hij zijn oordeel dan ook van een zodanige motivering te voorzien, dat deze voldoende inzicht geeft in de daaraan ten grondslag liggende gedachtegang om deze zowel voor partijen als voor derden, daaronder begrepen de hogere rechter, controleerbaar en aanvaardbaar te maken."[337]

Indien een van de partijen de deskundigheid in twijfel trekt, zal de rechter daar op moeten responderen.

4.2.10 Elektronische gegevens

Elektronische gegevens zijn geen geschriften, maar worden – indien zij met leestekens leesbaar gemaakt kunnen worden – wel op één lijn met geschriften gesteld. Hierbij kan gedacht worden aan de elektronische handtekening, elektronische akte of een overeenkomst die middels elektronische weg tot stand is gekomen. In het geval van de digitale handtekening, heeft de EU-Verordening 910/2014 (eIDAS-Vo) in artikel 3 onder 10, 11 en 12 drie categorieën digitale handtekeningen onderscheiden:

- 1. Elektronische handtekening: dit wordt ook wel de 'gewone' handtekening genoemd. Dit kunnen bijvoorbeeld handtekeningen zijn die zijn ingescand.
- 2. Geavanceerde elektronische handtekening: een elektronische handtekening, die voldoet aan de eisen van artikel 26:

  a. Zij is op unieke wijze aan de ondertekenaar verbonden;

  b. Zij maakt het mogelijk de ondertekenaar te identificeren;

  c. Zij komt tot stand met gegevens voor het aanmaken van elektronische handtekeningen die de ondertekenaar, met een hoog vertrouwensniveau, onder zijn uitsluitende controle kan gebruiken, en

  d. Zij is op zodanige wijze aan de daarmee ondertekende gegevens verbonden, dat elke wijziging achteraf van de gegevens kan worden opgespoord.

3. Gekwalificeerde elektronische handtekening: *"een geavanceerde elektronische handtekening die is aangemaakt met een gekwalificeerd middel voor het aanmaken van elektronische handtekeningen en die gebaseerd is op een gekwalificeerd certificaat voor elektronische handtekeningen"*[338]

De betrouwbaarheid van een handtekening wordt hier ook aan gekoppeld, een gekwalificeerde elektronische handtekening wordt betrouwbaarder dan een elektronische handtekening beschouwd.

4.2.11 Toepassing op deepfakes

Op basis van bovenstaande kan worden gesteld dat de betrouwbaarheid van een bewijsmiddel in beginsel wordt aangenomen, tenzij een van de partijen dit betwist. De rechter zal dus uitgaan van de authenticiteit van een bewijsmiddel en zelf niet op voorhand onderzoek doen naar de betrouwbaarheid, enkel als een van de partijen hierom verzoekt. Bovendien geldt op grond van artikel 21 Rv een waarheidsplicht, dit kan als een waarborg dienen dat ingebrachte bewijs daadwerkelijk authentiek is. Wanneer de waarheidsplicht wordt geschonden, kan de rechter de gevolgtrekking eraan verbinden die hij geraden acht.[339] Rechters hebben tevens veel vrijheid bij de sanctionering van een schending van de waarheidsplicht.

In geval van een getuigenverklaring is het belangrijk dat een rechter door blijft vragen om de authenticiteit vast te kunnen stellen. Dit is in het geval van een deepfake wat lastig, wel zou een rechter de eerste drie gestelde criteria als uitgangspunt kunnen gebruiken bij de beoordeling van de betrouwbaarheid van het aangeleverde bewijs, of er al dan niet sprake is van een deepfake. Allereerst zou dus gekeken kunnen worden naar de relevantie van het bewijsmiddel, daarnaast of het bewijsmiddel overeenkomt met andere bewijsmiddelen. Ook de kwaliteit kan getoetst worden; lijken de beelden waarheidsgetrouw of is er op enig moment een onjuistheid te zien, zoals het ontbreken van knipperende ogen?

Meer voor de hand liggend is dat een rechter een deskundige zal inschakelen indien de authenticiteit van een bewijsstuk wordt betwist, doordat er wordt gesteld dat er sprake is van een deepfake. Of in het geval dat juist wordt betwist dat er sprake is van een deepfake. Indien beide partijen het over eens zijn dat een deskundige wordt ingeschakeld, zou dat betekenen dat de uitkomst van het rapport leidend is ten aanzien van de authenticiteit van het bewijsmiddel. Een rechter heeft in het geval van een deskundige niet de plicht zich ruimschoots te verwoorden in zijn motivering waarom hij de deskundige volgt in zijn bevindingen.

De standaardregel van artikel 150 Rv geeft alle handvatten aan een rechter, omdat het de verdeling van de bewijslast en de daaropvolgende uitzonderingen bevat. Deze hoofdregel zal het uitgangspunt blijven, ook met betrekking tot deepfakes. Een rechter wordt namelijk vaker geconfronteerd met vals bewijs, zoals een vals opgemaakte akte of een valse handtekening.

De Hoge Raad heeft in 1993 in zijn arrest besloten dat in het geval van een betwisting van de echtheid van de tekst van een onderhandse akte, de bewijslast aan de hand van artikel 177 Rv (thans artikel 150 Rv) zal moeten worden verdeeld. Dit brengt als hoofdregel mee dat degene die stelt dat de akte is vervalst, de bewijslast zal dragen en daardoor ook het bewijsrisico heeft. Wel

stelt de Hoge Raad dat de rechter op grond van vaststaande feiten, zoals onverklaard gebleven onregelmatigheden in de tekst, of op grond van de onwaarschijnlijkheid van de stellingen van degene die de akte inroept, met betrekking tot de totstandkoming van de tekst tot het oordeel komen dat, behoudens tegenbewijs, moet worden aangenomen dat de tekst geheel of ten dele later boven de handtekening is geplaatst.[340] Tevens kunnen de vermelde eisen van redelijkheid en billijkheid van artikel 177 Rv meebrengen dat de bewijslast dient te rusten op degene die een beroep op de akte doet. Hiertoe heeft de rechter wel een motiveringsplicht.[341] Hier is in 2000 door de Hoge Raad aan toegevoegd dat de rechter hierbij alle omstandigheden die het in dit verband van belang achtte mag laten meewegen, waaronder ook de naar zijn oordeel onwaarschijnlijkheid van de stellingen van de eiser.[342] De rechter heeft dus een grote vrijheid bij de waardering van het bewijs.

Artikel 150 Rv zal bij deepfakes ook het uitgangspunt blijven. Een rechter zal in de basis uitgaan van de echtheid van het ingebrachte bewijsstuk, tenzij wordt gesteld dat dit vals is. De partij die dit stelt, zal dan uiteindelijk de bewijslast dragen en daardoor ook het bewijsrisico hebben dat hieraan is gekoppeld. Wanneer op basis van bijvoorbeeld onregelmatigheden in de video – bijvoorbeeld het ontbreken van knipperende ogen – kan de rechter aannemen dat er sprake is van een deepfake, behoudens tegenbewijs dat er geen sprake is van een deepfake. De rechter zal dit besluit dan wel moeten motiveren.

4.2.12 Artikel 6:162 BW

Op meerdere plekken in dit rapport is ingegaan op de onrechtmatige daad in combinatie met deepfakes. Om een zo volledig mogelijk beeld te kunnen schetsen, zal in dit hoofdstuk ook worden ingegaan op het bewijsrecht rondom de onrechtmatige daad.

Op grond van artikel 6:162 BW kan degene die stelt schade te hebben geleden door een onrechtmatige daad van een ander, deze schade proberen te verhalen op deze persoon middels een gerechtelijke procedure. Allereerst zal deze partij moeten stellen en motiveren dat aan de vereisten van artikel 6:162 BW is voldaan. Aan de materiele vereisten zal moeten worden voldaan. Er dient sprake te zijn van een 'daad/gedraging', 'onrechtmatigheid', 'schade' welke 'toerekenbaar' moet zijn en tussen de schade en de gedraging moet een 'causaal verband' aanwezig zijn. Tot slot dient er sprake te zijn van 'relativiteit', dit is uitgewerkt in artikel 6:163 BW. Indien de ene partij dit stelt, zal de rechter op grond van artikel 24 Rv moeten beslissen over deze onrechtmatigheid. Hierbij zal de eiser van de vordering voldoende op het verweer van de gedaagde moeten ingaan, als de eiser dat niet doet kan de rechter de vordering namelijk afwijzen.

De rechter beslist op de grondslag van hetgeen is gebleken en is vast komen te staan. De rechter is overigens niet bevoegd om uit eigen beweging feiten aan zijn beslissing ten grondslag te leggen, het dient als ware hem letterlijk aangereikt te worden. Hij dient zich dus terughoudend op te stellen. Op grond van artikel 149 lid 2 Rv geldt dit niet voor feiten of omstandigheden van algemene bekendheid en ervaringsregels. In het geval van feiten of omstandigheden van algemene bekendheid of ervaringsregels, hoeven deze niet worden voorgelegd door een van de partijen maar kan de rechter deze dus zelf aan zijn beslissing ten grondslag leggen. De beslissing van de rechter dat een

feit of een bepaalde omstandigheid een feit of omstandigheid van algemene bekendheid is, is een feitelijke beslissing die in cassatie beperkt getoetst kan worden.[343] Verder is de rechter vrij zich ter ondersteuning van de feitelijke grondslag ook op andere feiten te laten berusten, mits hij hierbij wel binnen de grenzen van artikel 149 Rv blijft.

4.2.13 Conclusie

Een onrechtmatige daad op het gebied van een deepfake, zou betrekking kunnen hebben op de privacy en de goede naam van de persoon in kwestie, waarbij de persoon schade heeft geleden door het vervaardigen en/of verspreiden van de deepfake. Bescherming van de eer en goede naam van een persoon kan middels een onrechtmatige daad-procedure worden bewerkstelligd. Vanuit dit oogpunt zal naar de stelplicht- en bewijslastverdeling worden gekeken. Gemakshalve zal worden uitgegaan van voldane vereisten, waarbij alleen de schade nog onduidelijk is.

Indien de eiser stelt schade te hebben geleden, is de schade onderwerp van de zaak. Indien een van de vereisten van artikel 6:162 BW onderwerp is van de rechtszaak, zal de rechter op grond van artikel 24 Rv hierover beslissen. Dit is dus ook het geval als de zaak over de geleden schade gaat. Op grond van artikel 6:97 BW zal de rechter de schade begroten. Hierbij krijgt de rechter een grote mate van vrijheid, waarbij de rechter niet gebonden is aan de regels van stelplicht en bewijslast.[344] De rechter is hierbij vrij om, wanneer feiten zijn vast komen te staan waaruit de geleden schade opgemaakt kan worden, gerechtigd de schade zelf te begroten.[345] Wel heeft de Hoge Raad in 2009 geoordeeld dat hoewel het een rechter vrij staat om de schade te begroten en hierbij van de gewone regels van de stelplicht en bewijslast af te wijken, belet dit hem geenszins de gewone regels van stelplicht en bewijslast toe te passen bij een geschil over de feiten die in de discussie over de schadeomvang worden gesteld en waarvan de rechter acht dat deze belangrijk zijn voor de schadebegroting.[346]

Het kan ook voorkomen dat de eiser niets zegt over de schade die hij heeft geleden. In 1991 heeft de Hoge Raad hierover beslist dat voor toewijzing van een vordering tot schadevergoeding het voldoende is dat de feiten zijn gesteld en komen vast te staan waaruit in het algemeen de schade kan worden afgeleid. De rechter is dan vrij om, met in achtneming van de aard van de schade, zonder nader bewijs aannemelijk te achten dat de schade is geleden en vrij de schade vervolgens in te schatten.[347] Belangrijk is hierbij op te merken dat voor toewijzing van de schade op z'n minst feiten moeten zijn gesteld, waarop de schade is in te schatten, maar de rechter vrij is in het toewijzen van de schade en het schatten van de hoogte van de schade.

4.3 Strafprocesrecht

4.3.1 Algemeen

Het bewijsrecht moet de waarborgen voor adequate bewijsbeslissingen bieden. Nadat de rechter de formele vragen van artikel 348 Sv langs is gelopen en er geen reden tot schorsing van de vervolging aanwezig is, zal de rechter de eerste materiële hoofdvraag moeten beantwoorden: is bewezen dat het tenlastegelegde feit door de verdachte is begaan?[348] Bewijzen kan worden gezien als het aantonen dat in redelijkheid niet kan worden getwijfeld aan de juistheid van het verwijt dat aan de verdachte wordt gemaakt.

Nederland kent een negatief-wettelijk bewijsstelsel. Dit houdt onder meer in dat er sprake is van bewijsminima, waarbij de rechterlijke overtuiging een belangrijke rol speelt. Op de verdachte rust geen bewijslast – deze rust daarentegen op de officier van justitie – en zodoende ook geen bewijsrisico. Het bewijsrisico houdt in dat een veroordeling uitblijft omdat de rechter van oordeel is dat wettig en overtuigend bewijs ontbreekt. Dit risico realiseert zich als gevolg van het niet nakomen van de bewijslast.[349] Als een verdachte wel een bewijslast zou hebben, zou dit onverenigbaar zijn met de onschuldpresumptie. De officier van justitie zal de feiten en omstandigheden aandragen om zo de door hem opgestelde tenlastelegging te ondersteunen.[350]

Negatief-wettelijk betekent dat het bewijs alleen met in de wet opgesomde bewijsmiddelen mag worden geleverd. Negatief duidt erop dat de rechter het feit niet bewezen mag verklaren, als hij vanuit de bewijsmiddelen niet tot de overtuiging is gekomen dat de verdachte het ten laste gelegde heeft begaan. Het bewijsstelsel is wettelijk, omdat de wet de bewijsmiddelen limitatief opsomt. Als tweede kenmerk geldt dat het Nederlandse stelsel enkele bewijsminimumregels bevat. De rechter is voor het bewijs niet aan kwantitatieve maatstaven gebonden, wel geldt de regel dat het tenlastegelegde feit niet kan worden aangenomen op enkel de bekentenis van de verdachte of enkel de verklaring van een getuige.[351]

Het Nederlandse bewijsstelsel wordt gekenmerkt door de vrijheid van waardering die de rechter toekomt. De grondslag van het bewijsrecht is neergelegd in artikel 338 Sv.[352] Dit artikel zegt dat de rechter het bewijs aan mag nemen als hij de overtuiging heeft bekomen door de inhoud van wettige bewijsmiddelen. Overtuiging houdt 'een zeer klemmende graad van waarschijnlijkheid' in, en moet op bewijsmiddelen zijn gebaseerd.

Alle bewijsmateriaal dat via de wettige bewijsmiddelen wordt aangedragen, mag in zowel de procedure als in de bewijsbeslissing worden gebruikt, behalve in gevallen dat het bewijsmateriaal wordt uitgesloten om redenen van (on)betrouwbaarheid dan wel (on) rechtmatigheid.[353]

4.3.2 Feiten of omstandigheden van algemene bekendheid

Ook geldt dat feiten of omstandigheden van algemene bekendheid geen bewijs behoeven; wel kunnen deze in samenhang met ander bewijsmateriaal duiding geven. Dit zijn gegevens *"die ieder van de rechtstreeks bij het geding betrokkenen geacht moet worden te kennen of die hij zonder noemenswaardige moeite uit algemeen toegankelijke bronnen kan achterhalen. Voor algemene ervaringsregels geldt hetzelfde".*[354] De Hoge Raad lijkt met 'ieder van de rechtstreeks bij het geding betrokkenen' een ruimere interpretatie voor ogen te hebben dan alleen de procesdeelnemers.[355] In het geval van feiten of omstandigheden van algemene bekendheid gaat het voornamelijk om feiten en omstandigheden waarvoor geen specialistische kennis nodig is. Het kan bijvoorbeeld gaan om historische gebeurtenissen, zoals het lot van de joden in de Tweede Wereldoorlog, de aanslagen op 9/11, maar ook andersoortige gebeurtenissen zoals dat laptops veelvuldig onderwerp van diefstal zijn,[356] of dat Schiphol wordt gebruikt voor de in-, uit- en doorvoer van voorwerpen die onmiddellijk of middellijk afkomstig zijn uit een misdrijf.[357] Het gevaar bij feiten en omstandigheden van algemene bekendheid is dat rechters een gegeven

te snel als zodanig beschouwen. Het heeft namelijk als gevolg dat ter zake daarvan geen bewijsmiddelen in het vonnis te hoeven worden opgenomen. Een feit van algemene bekendheid behoeft namelijk geen bewijs en hoeft ook niet ter sprake te zijn gebracht op de zitting. Daardoor is een rechter wel verplicht, indien een discussie is ontstaan over de vraag of iets aangemerkt kan worden als algemene bekendheid, dit wel ter orde te stellen tijdens het onderzoek ter terechtzitting.

4.3.3 Wettige bewijsmiddelen

De bewezenverklaring dient te zijn gefundeerd op basis van wettige bewijsmiddelen. Deze hoofdregel is uitgewerkt in artikel 339 lid 1 en 340 tot en met 344 Sv. Onder wettige bewijsmiddelen vallen de eigen waarneming van de rechter, verklaringen van de verdachte, een getuige of een deskundige en schriftelijke bescheiden. In het kader van dit onderzoek zal er voornamelijk aandacht worden besteed aan de eigen waarneming van de rechter, zoals dat is neergelegd in artikel 339 lid 1 en 340 Sv. Via de eigen waarneming van de rechter kunnen bronnen van bewijs worden geïntroduceerd, die ten tijde van het ontstaan van het Wetboek van Strafvordering nog niet bestonden, zoals bewijs in de vorm van video- en geluidsopnamen. Ten aanzien van de eigen waarneming van de rechter zijn in de wet geen bewijsminimumregel opgenomen. Wel wordt door de meeste auteurs in de literatuur uitgegaan van een algemene eis van dubbele bevestiging.[358]

De beelden die als bewijs moeten dienen kunnen aan de rechter worden getoond, zodat hij vervolgens hetgeen hij zag tot bewijs kan bezigen. Bij het waarnemen moet worden gedacht aan het middels de zintuigen registreren. De waarneming van de rechter moet persoonlijk zijn gedaan; hierbij geldt wel dat de rechter in hoger beroep gebruik kan maken van de eigen waarneming van de rechter in eerste aanleg.[359] De rechter is vrij te kiezen of hij de waarnemingen die hij doet ter terechtzitting, ter sprake brengt, tenzij het een verrassing voor de procespartijen oplevert.[360] De eigen waarneming dient alleen ter sprake te worden gebracht *'indien de procespartijen door het gebruik van die waarneming voor het bewijs zouden worden verrast omdat zij daarmee geen rekening behoefden te houden. Of daarvan sprake is, is afhankelijk van de omstandigheden van het geval, zoals het procesverloop, de aard van de waarneming en het verband van die waarneming met het voorhanden bewijsmateriaal'.*[361]

Onder de eigen waarneming kunnen dus bewijs in de vorm van foto-, film-, audio- en videomateriaal vallen, maar ook tekeningen, schetsen, diagrammen et cetera. Ook kunnen zichtbare fysieke kenmerken van bijvoorbeeld de verdachte worden waargenomen door de rechter. De rechter kan zodoende iemands uiterlijk vaststellen. Het horen van een verdachte, getuige of deskundige tijdens het onderzoek ter terechtzitting valt niet onder de eigen waarneming. Daarnaast moet de waarneming ter terechtzitting zijn gedaan, zo wordt namelijk voorkomen dat procespartijen en derden op een later moment worden geconfronteerd met resultaten van waarnemingen van een rechter die buiten hen om hebben plaatsgevonden. Dit is namelijk in strijd met de interne openbaarheid. Het is overigens niet vereist dat de waarneming in de zaak tegen de verdachte is gedaan, dit betekent dat de waarneming ook tijdens het onderzoek ter terechtzitting tegen een medeverdachte mag worden gedaan. Uitzondering op het vereiste dat de waarneming ter terechtzitting moet zijn gedaan, zijn de foto's en videobanden die

gevoegd zijn bij het dossier.[362] Dit heeft de Hoge Raad in zijn arrest van 24 september 2019 ook geoordeeld.[363]

4.3.4 Valse bewijsstukken

Het strafrecht is al enige tijd bekend met onware bewijsstukken, zoals bijvoorbeeld valse bekentenissen. Dat deze valse bekentenissen grote gevolgen hebben, is onder meer duidelijk geworden in zaken als Central Park jogger case uit 1989, waarbij een jonge vrouw werd verkracht en bijna werd vermoord toen zij in het Central Park hardliep. De politie bracht vervolgens vijf zwarte en hispanic jongens van tussen de veertien en zestien naar het bureau voor het onderzoek, waarna alle vijf na zo'n dertig uur te zijn ondervraagd, bekenden iets te maken hebben gehad met deze verkrachting. Achteraf blijkt dat zij het niet hebben begaan, maar door de ondervragingstechnieken hebben zij deze bekentenissen toch gedaan. Valse bekentenissen komen ook voor in het Nederlandse rechtssysteem, zoals in het geval van de Schiedammer Parkmoord,[364] de Puttense moordzaak,[365] de zaak van de bejaardenverzorgster Ina Post,[366] de Arnhemse villamoord.[367] Naast valse bekentenissen kunnen natuurlijk ook de inhoud van getuigenverklaringen, processen-verbaal en schriftelijke bescheiden onwaarheden bevatten. Dat kan opzettelijk zijn, maar ook onbewust en onbedoeld. Getuigenverklaringen zijn lang niet altijd betrouwbaar, ook als getuigen naar eer en geweten rapporteren wat zij denken te hebben gezien. Schriftelijke bescheiden kunnen vervalst zijn, maar ook simpelweg fouten of onwaarheden bevatten. Dat geldt zeker ook voor digitaal bewijs, dat afgedrukt (bijvoorbeeld een print van een emailbericht) als schriftelijk bescheid, in de vorm van een deskundigenverklaring van een forensisch analist, of als gevoegd bij de processtukken kan worden ingebracht in de rechtszaak.

4.3.5 Bewijsrecht in verhouding tot Deepfakes

In het huidige Wetboek van Strafvordering is geen specifiek artikel opgenomen dat het bewijsrecht rondom deepfakes reguleert. Dit is ten aanzien van bewijs waarvan aangetoond kan worden dat het daadwerkelijk een deepfake betreft niet problematisch. Wel is dit problematisch ten aanzien van bewijs waarvan de authenticiteit bevestigd noch ontkend kan worden. Uit hoofdstuk twee is namelijk gebleken dat ongeveer 65% van de deepfakes wordt gedetecteerd. Omdat er geen specifiek artikel gericht op deepfakes bestaat, zal de bestaande doctrine omtrent bewijs in strafzaken, ook hier toepassing vinden. Echter, dit roept de vraag op of het bestaande bewijsrecht in het Wetboek van Strafvordering toereikend is ten aanzien van bewijs waarvan de authenticiteit niet zeker is.

4.3.6 Wijze waarop bewijs wordt gewaardeerd

De wijze waarop bewijs wordt gewaardeerd door rechters is onderhevig aan veel onduidelijkheid. De motiveringsplicht bij een vonnis is allereerst al neergelegd in artikel 121 van de Grondwet, waarin wordt bepaald dat vonnissen de gronden inhouden waarop zij berusten. Dit wordt herhaald in artikel 359 Sv. Meer specifiek wordt in artikel 359 lid 3 Sv bepaald dat in geval van een bewezenverklaring het vonnis de inhoud van de bewijsmiddelen dient te bevatten waarop de beslissing berust. Daarnaast moet de inhoud van die bewijsmiddelen redengevend zijn voor de beslissing en moet het gebruik van bepaalde bewijsmiddelen nader worden gemotiveerd, op grond van artikel 360 lid 1 en 2 Sv.

De legitimatie van het strafrecht is afhankelijk van meerdere waarborgen waarmee het

bewijsoordeel in strafzaken is omgeven. Een van de waarborgen is de plicht van de rechter om de bewijsbeslissing te motiveren. De motivering en de normering van de bewijsvoering en bewijswaardering zijn onlosmakelijk met elkaar verbonden.[368] Hiervoor heeft de rechter een ruime beoordelingsvrijheid, waarmee hij tot de 'overtuiging' komt, zoals dat in artikel 338 Sv wordt beschreven. Door Corstens worden drie functies van de motiveringsverplichting onderscheiden.[369]

Allereerst wordt de inscherpingsfunctie genoemd. De motivering vormt een aanvulling op de normering van het bewijsoordeel, wat in de literatuur als de inscherpingsfunctie van de motivering wordt genoemd. Dit houdt in dat omdat de rechter verplicht wordt te motiveren, hij gedwongen wordt tegenover zichzelf rekenschap af te leggen van zijn motieven. Een rechter zal na het lezen van het dossier een voorlopig oordeel over de zaak hebben, door een rechter te verplichten zijn oordeel te motiveren, wordt hij *"gedwongen zichzelf rekenschap te geven van zijn beweegredenen"*.[370] De rechter wordt zodoende scherper met de vragen geconfronteerd en kan dus niet zomaar volstaan met formuleren van conclusies. Als tweede functie wordt de explicatiefunctie genoemd. Door te motiveren licht de rechter de procespartijen en derden in omtrent de gronden die hij aanvoert en zo worden de partijen ingelicht op welke gronden de beslissing steunt. Wel dient hierbij een kanttekening gemaakt te worden, de opgegeven motivering behoeft namelijk niet overeen te stemmen met de motieven die de rechter hebben bewogen. Dit geldt met name in zaken die voor een meervoudige kamer zijn gebracht, omdat de opvattingen van de rechters uiteen kunnen lopen. De laatste functie is de controlefunctie: motivering maakt de toetsing door bijvoorbeeld de appel- of cassatierechter gemakkelijker. Indien de rechter in eerste aanleg zijn oordeel onvoldoende motiveert, is het voor de appèlrechter lastiger om zijn eigen oordeel in de plaats van die van de rechter in eerste aanleg te stellen. Voor de cassatierechter zal het bovendien moeilijker worden te beoordelen of de gedane beslissing door de aangevoerde gronden kan worden gedragen. De toetsing van een uitspraak in hoger beroep en cassatie wordt dus vergemakkelijkt wanneer een oordeel goed is gemotiveerd.

4.3.7 Verweren

Bewijsverweren zijn op te delen in verweren die zich richten op de totstandkoming van het bewijsmateriaal, verweren tegen de inhoud van het bewijsmateriaal en verweren die zich richten op de aansluiting tussen het bewijsmateriaal en de tenlastelegging. Bovendien zijn er ook mengvormen van deze drie categorieën.

Bewijsverweren kunnen zich richten tegen de wettigheid van het bewijs, bijvoorbeeld als een opsporingsambtenaar niet bevoegd was het proces-verbaal op te stellen. Anderzijds kan het gericht zijn tegen de rechtmatigheid, zoals het onrechtmatig verkrijgen van bewijs, verder kan het gericht zijn op de 'redengevendheid', wat inhoudt dat wanneer bewijsmiddelen niets inhouden dat werkelijk van betekenis is daartegen een bewijsverweer kan worden aangevoerd. Ook kan het betrekking hebben op de 'toereikendheid' van de bewijsmiddelen, waarin wordt betwist of het bewijsmateriaal toereikend is om de bewezenverklaring te kunnen ondersteunen. Als laatste wordt het bewijsverweer tegen de betrouwbaarheid van het bewijs, hier zal vanwege de relevantie voor dit rapport dieper op worden ingegaan.

4.3.8 Bewijsverweer tegen de betrouwbaarheid

Betrouwbaarheidsverweren richten zich tegen de betrouwbaarheid, deugdelijkheid of geloofwaardigheid van de inhoud van de bewijsmiddelen. De rechter is bevoegd ten aanzien van de selectie en waardering van de inhoud van een bewijsmiddel. De rechter heeft hiertoe het laatste woord. De rechter is bovendien bevoegd te beslissen welk bewijs wordt meegenomen in de beslissing en aan welk bewijs geen waarde wordt gehecht. De verdediging heeft wel de kans om zich uit te laten over de inhoud van de aangedragen bewijsmiddelen. Indien de verdediging dit doet, zal de rechter op grond van artikel 359 lid 2 Sv een verweer gemotiveerd moeten weerleggen.[371] De Hoge Raad geeft niet aan wat betrouwbaarheid precies inhoudt of hoe de rechter de betrouwbaarheid of onbetrouwbaarheid vast moet stellen. Dit geldt eveneens ten aanzien van het betrouwbaarheidsverweer, hier worden namelijk ook niet veel eisen gesteld.

Het is onduidelijk hoe rechters een bewijsmiddel op betrouwbaarheid onderzoeken, wel blijkt uit de praktijk dat zij op zoek zijn naar het waarheidsgehalte van de inhoud van het bewijsmiddel. Althans, in geval van bewijsmiddelen in de vorm van getuigenverklaringen. In dat geval wordt vooral gekeken naar de consistentie van verklaringen. Desondanks zijn er geen uitdrukkelijke criteria geformuleerd waaraan de betrouwbaarheid kan worden getoetst. De rechter is zodoende vrij in de waardering van de betrouwbaarheid en de bewijswaarde die hij toekent aan de verklaring.

Ten aanzien van bewijs in de vorm van een videofragment of een foto wordt dit bewijs geschaard onder de eigen waarneming van de rechter. Deze zal zich moeten realiseren dat de authenticiteit van beeldmateriaal niet altijd zeker is. Dat is op zich al lang het geval; foto's en video's kunnen immers bewerkt zijn. Toch heeft beeldmateriaal, zo vermoeden wij, nog vaak een aura van authenticiteit, omdat foto's en films van oudsher de werkelijkheid afbeelden (in de zin dat zij een echte weergave bieden van een situatie ter plekke; die situatie kan natuurlijk, zoals bij een filmset, fictief zijn). Niettemin is sinds enkele decennia genoegzaam bekend dat men met speciale effecten en foto- en videoprogramma's (het spreekwoordelijke 'fotoshoppen') situaties kan tonen die zich niet in de werkelijkheid hebben voorgedaan of die de werkelijke situatie in sommige opzichten vertekenen. In die zin vormt deepfaketechnologie geen nieuw fenomeen. Wel zal de schaal waarop en de mate waarin aan de authenticiteit van beeldmateriaal kan of moet worden getwijfeld, door deepfakes gaan toenemen.

De vraag of een bron al dan niet authentiek is, zal normaliter pas ter sprake komen indien een van de partijen dit betwist. Daarna zal de rechter de betrouwbaarheid beoordelen. Als echter deepfakes een substantieel aandeel gaan krijgen in de totale hoeveelheid beeldmateriaal dat rondgaat, zal de rechter ook uit zichzelf moeten afvragen of een video authentiek en betrouwbaar is. Mogelijk zouden in zo'n situatie ook het Openbaar Ministerie en de verdediging standaard een forensische authenticiteitscheck moeten (doen) uitvoeren voor al het beeldmateriaal dat als mogelijk bewijs wordt ingebracht, zodat elke video bijvoorbeeld een betrouwbaarheidslabel (vrijwel zeker authentiek / vermoedelijk authentiek / onduidelijk / vermoedelijk niet authentiek / vrijwel zeker niet authentiek) meekrijgt. Voor

zover dat niet gebeurt, bestaat de mogelijkheid dat partijen in de rechtszaal de betrouwbaarheid van elke ingebrachte video gaan betwisten, hetgeen de nodige verzwaring van de rol van forensisch deskundigen zal betekenen, die het kat-en-muisspel tussen deepfaketechnologie en deepfakedetectietechnologie nauwgezet zullen moeten volgen. Aangezien deze technologie, en meer in het algemeen kunstmatige intelligentie waarvan deepfaketechnologie gebruik maakt, nog een jonge tak van wetenschap is, is daarbij een vraag of rechters voldoende handvatten hebben om de deskundigheid van deepfake-deskundigen te beoordelen. Potentieel kunnen deepfakes daarom een ingrijpende impact hebben op de rechtspraktijk waar het gaat om de beoordeling van het bewijs. Tegelijkertijd kan dit risico ook enigszins worden gerelativeerd, aangezien bij de opkomst van digitale technologie vergelijkbare zorgen opkwamen vanwege de eenvoudige manipuleerbaarheid van digitale gegevens; dit heeft echter niet geleid tot fundamentele problemen in bewijsrecht of praktijk.

4.3.9 Conclusie

De mogelijke disruptieve effecten van de introductie van deepfakes in de strafrechtszaal zijn groot. Zoals in hoofdstuk 2 duidelijk werd kan het strafrecht onder druk komen te staan door ten minste een viertal punten. Ten eerste kunnen processen langer duren en onzekerder worden, omdat partijen altijd kunnen beweren dat tegen hen geleverd bewijs nep en gefabriceerd is. Ten tweede is het gevaar dat de rechter inhoud onterecht voor waar zal aannemen en dit tot een onterechte veroordeling leidt. Ten derde kan een veroordeelde, na een rechtelijke uitspraak, altijd publiekelijk zijn onschuld volhouden door te beweren dat de rechter in een fake-bericht is gestonken. Ten vierde kan bij bepaalde delicten de suggestie die een deepfake oplevert al genoeg zijn voor een publieke veroordeling. Zowel in de literatuur als in de voor deze studie gehouden interviews (zie paragraaf 6.4) wordt dit probleem veelvuldig genoemd. De laatste twee mogelijke effecten zijn belangrijke maatschappelijke risico's; de eerste twee betreffen wezenlijke juridische risico's, die onzes inziens zeker nader onderzoek verdienen.

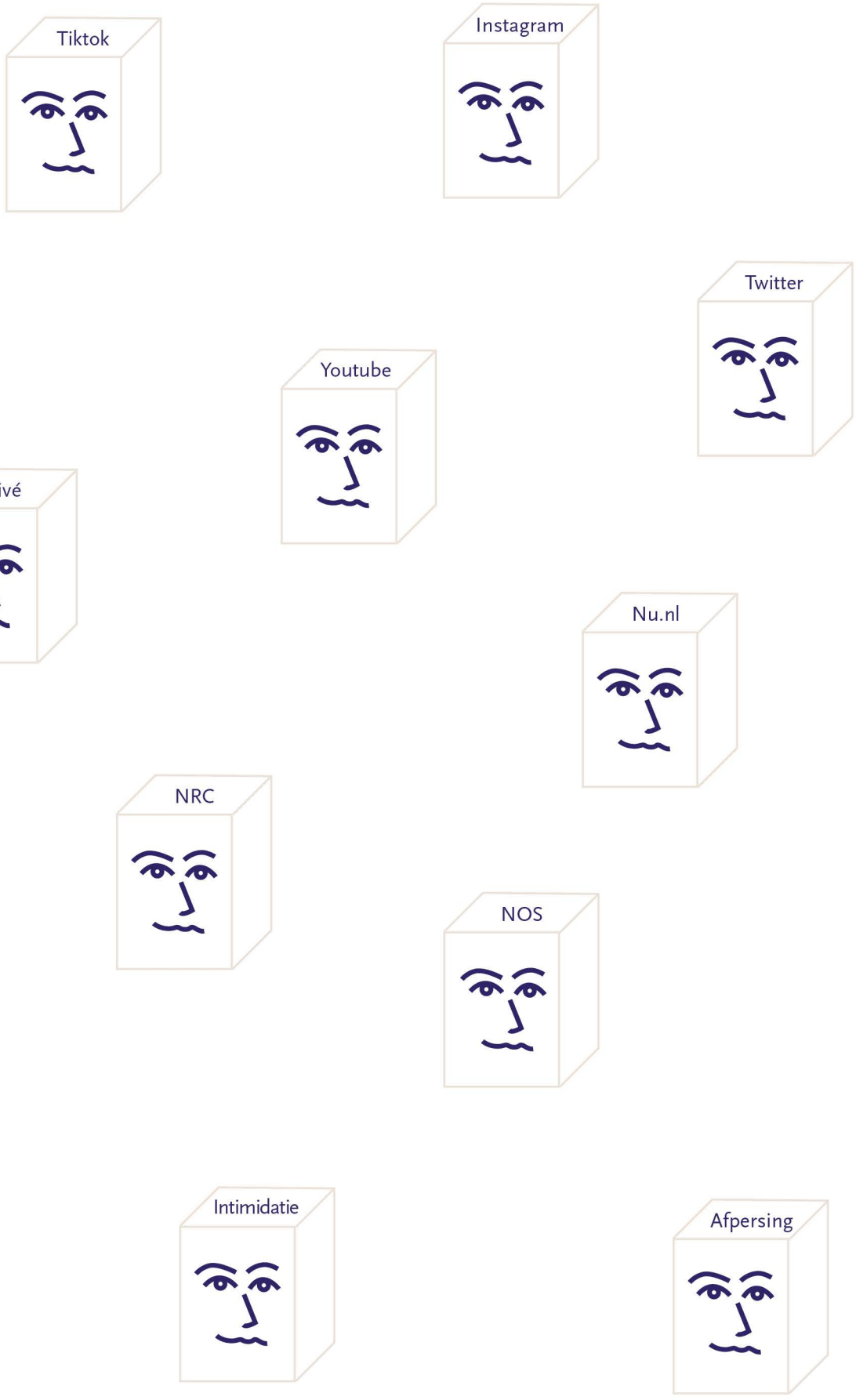
4.4 Conclusie

Uit dit hoofdstuk blijkt dat er zowel binnen het civielrecht als binnen het strafrecht een complex stelsel aan regels, indicaties en contra-indicaties speelt bij mogelijke bewijsvraagstukken in het kader van deepfake video's, afbeeldingen, audiofragmenten of ander materiaal. Zowel het burgerlijk procesrecht als het strafprocesrecht zijn redelijk open van aard. Er bestaan geen bijzondere bepalingen ten aanzien van deepfakes. Net zoals bij de materieelrechtelijke bepalingen, zijn de procesrechtelijke regels op zich breed genoeg om vraagstukken omtrent de authenticiteit van deepfakes te adresseren. In die zin zijn deepfakes slechts de zoveelste variant van technische mogelijkheden om bewijsmateriaal te manipuleren of te fabriceren. Toch geldt ook hier dat zowel de ogenschijnlijke echtheid en de toegang tot dergelijke middelen door iedere burger een wezenlijk risico vormen voor juridische processen.

Voetnoten hoofdstuk 4

- 308 Pfefferkorn, R. (2020). "Deepfakes" in the courtroom. *Boston University Public Interest Law Journal, 29*(2).
- 309 Roell, T. (2008). 'Bewijslastverdeling voor beginners',Advocatenblad, p. 498.
- 310 T&C Rv, commentaar op art. 150 Rv
- 311 Roell, T. (2008). 'Bewijslastverdeling voor beginners',Advocatenblad, p. 499.
- 312 HR 09-09-2005, ECLI:NL:PHR:2005:AT5156
- 313 ECLI:NL:HR:2020:1510
- 314 ECLI:NL:HR:2020:1510, r.o. 3.4.
- 315 T&C, artikel 150 Rv.
- 316 ECLI:NL:GHSHE:2019:1337
- 317 Asser Procesrecht 3 Bewijs, 56 Tegenbewijslevering.
- 318 HR 05-04-2013, ECLI:NL:HR:2013:BY8101 (Lundiform/Mexx)
- 319 HR 12 januari 2001, NJ 2001/419 (J./ABN Amro c.s.)
- 320 HR 17 april 2009, LJN: BH2955, NJ 2009/196
- 321 Van Mierlo, T&C Rv, commentaar op artikel 20 Rv.
- 322 HR 18-03-2011, ECLI:NL:HR:2011:BP0571, m.nt. C.J.M. Klaassen (Antebi/Joodse Gemeente)
- 323 Nederlands Juristenblad, Modernisering civiel bewijsrecht
- 324 Van Mierlo, T&C Rv, commentaar op artikel 22b Rv.
- 325 Asser Procesrecht 2 Eerste aanleg, 98 Verboden aanvulling van of met feiten en rechten.
- 326 Asser Procesrecht/Asser 3 2017/152, 'Informatie voor het bewijs'.
- 327 Asser Procesrecht/Asser 3 2017/247.
- 328 Asser Procesrecht/Asser 3 2017/248.
- 329 Asser Procesrecht/Asser 3 2017/2548a.
- 330 Asser Procesrecht/Asser 3 2017/250, ' De echtheid en valsheid van akten; het bewijsrisico'.
- 331 Parlementaire Geschiedenis van de nieuwe regeling van het bewijsrecht in burgerlijke zaken, MvA TK (1981) [p. 153]
- 332 de Bock, R. H. Tussen waarheid en onzekerheid. *Over het vaststellen van feiten in de civiele procedure, 1. Hoofdstuk: Getuigenbewijs en waarheidsvinding.*
- 333 De Groot, G. &. Elbers, N.A. (2008).' Inschakeling van deskundigen in de rechtspraak', Raad voor de rechtspraak.
- 334 Van Dijk, Chr.H. (2007). ' Hoe te beoordelen of de deskundige deskundig is?' , NTBR, 61.
- 335 HR 16-10-1998, ECLI:NL:HR:1998:ZC2743
- 336 HR 9 december 2011, ECLI:NL:HR:2011:BT2921 *(X en Y / Stichting Flevoziekenhuis)*
- 337 HR 09-122011, ECLI:NL:HR:2011:BT2921 (Flevoziekenhuis), r.o. 3.4.5.
- 338 Artikel 3 onder 12 EU-Verordening 910/2014 (eIDAS-Vo)
- 339 HR 6 juli 2018, ECLI:NL:HR:2018:1103, r.o. 3.3.2
- 340 HR 15 januari 1993, ECLI:NL:HR:1993:ZC0827, r.o. 3.5.
- 341 HR 15 januari 1993, ECLI:NL:HR:1993:ZC0827, r.o. 3.5.
- 342 HR 14 januari 2000, ECLI:NL:HR:2000:AA4278, r.o. 3.4.
- 343 Beenders, T&C commentaar op 149 Rv.
- 344 HR 28 juni 1991, NJ 1991, 746; HR 15 november 1996, NJ 1998, 314.
- 345 HR 8 april 2005, NJ 2005, 371 (Van de Ven/Van de Ven), r.o. 3.4
- 346 HR 5 juni 2009, NJ 2009, 257, r.o. 3.3.2.
- 347 HR 28 juni 1991, ECLI:NL:HR:1991:ZC0304, r.o. 3.3.
- 348 Corstens, G.J.M (2020), Het Nederlands strafprocesrecht, Kluwer.
- 349 <https://research.vu.nl/ws/portalfiles/portal/2897457/AAe.bewijsrisico%2C+bewijslast.pdf>.
- 350 <https://www.narcis.nl/publication/RecordID/oai:tilburguniversity.edu:publications%2F-62603d3a-fa2e-4a04-bf54-117cbbee94ec>.
- 351 Het Nederlandse bewijsrecht in strafzaken, Preadvies van Mr. A.L. Melai, p. 254.
- 352 Het Nederlandse bewijsrecht in strafzaken, Preadvies van Mr. A.L. Melai, p. 253.
- 353 Nijboer, J.F. (2004). Legaliteit en het strafrechtelijk bewijsrecht; uitholling van het wettelijk bewijsstelsel in strafzaken?, Ars Aequi, Volume: 53.
- 354 HR 11 januari 2011, LJN: BP0291, onder 3.2.1.
- 355 HR 11 januari 2011, LJN: BP0291 m.nt. P.A.M. Mevis.

- 356 HR 13 mei 2003, Nj 2003, 460.
- 357 HR 27 september 2005, ECLI:NL:HR:2005:AT4094.
- 358 Dubelaar, M. J. (2014). *Betrouwbaar getuigenbewijs: totstandkoming en waardering van strafrechtelijke getuigenverklaringen in perspectief.* Leiden University.
- 359 HR 17 januari 2012, ECLI:NL:HR:2012:BU6056, Nj 2012 66.
- 360 HR 5 juli 2016, ECLI:NL:HR:2016:1405
- 361 HR 15 december 2009 ECLI:NL:HR:2009:BJ2831, r.o. 3.5.3 en HR 25 september 2012, ECLI:NL:HR:2012:BX4990, r.o. 3.6.
- 362 T&C Strafvordering bij artikel 340 Sv
- 363 HR 24 september 2019, ECLI:NL:HR:2019:1414 (Eigen waarneming van de rechter)
- 364 HR 7 september 2004, ECLI:NL:GHS-GR:2002:AE0013
- 365 HR 26-06-2001, ECLI:NL:HR:2001:AA9800, m.nt. T.M. Schalken
- 366 23 juni 2009, ECLI:NL:PHR:2009:BI1689
- 367 2 februari 2021 ECLI:NL:HR:2021:633
- 368 <https://cris.maastrichtuniversity.nl/ws/portalfiles/portal/1044743/guid-de76f774-1357-4f78-ae49-31d5dd800058-ASSET1.0.pdf>.
- 369 Corstens, G.J.M (2020), Het Nederlands strafprocesrecht, Kluwer.
- 370 Corstens, G.J.M (2020), Het Nederlands strafprocesrecht, Kluwer, p. 781.
- 371 Stevens, L. (2014). 'Bewijs waarderen' in Bewijs waarderen in het strafrecht, NJB (89), p. 2844.



Een deepfake kan de werkelijkheid imiteren, of een nieuwe werkelijkheid creëren, voor verschillende doeleinden.
Van onschuldig satire tot bewust een reputatie schaden of politieke onrust aanwakkeren.

5. Handhaving en toezicht

In dit hoofdstuk wordt kort ingegaan op het toezicht op en de handhaving van de materieelrechtelijke bepalingen. Eerst zal een algemene schets van deze problematiek worden geboden aan de hand van eerder op dit punt verschenen rapporten (paragraaf 5.1), dan wordt vervolgd met een tweetal specifieke vraagstukken, namelijk de keuze tussen *ex ante* en *ex post* regulering (paragraaf 5.2) en de keuze tussen de normering van verschillende spelers in de keten (paragraaf 5.3), vervolgens wordt kort stilgestaan bij de regels omtrent de verantwoordelijkheid en aansprakelijkheid van internet intermediairs als vervat in de Digital Services Act die momenteel in ontwikkeling is (paragraaf 5.4) en tot slot wordt afgesloten met een korte conclusie (paragraaf 5.5). Daarbij moet worden opgemerkt dat de grens tussen regulering *ex ante* en *ex post* niet eenduidig te trekken is; vaak zijn er grijstinten of elkaar deels overlappende reguleringsbenaderingen. Toch is het verschil van belang. De meeste regulering van technologie is momenteel *ex post*, dat wil zeggen dat er met name grenzen worden gesteld aan het gebruik van technologische middelen en niet aan de productie, de verkoop of de verspreiding van de technologische middelen. Daarbij komt dat er meestal pas wordt geverifieerd of bepaald gebruik binnen de wettelijke grenzen is gebleven nadat het gebruik reeds plaatsgegrepen heeft. Er kan echter ook eerder in het proces worden ingegrepen, ofwel bij de bron, namelijk bij de productie, verkoop of verspreiding van technologische middelen, ofwel voordat deze middelen worden gebruikt of ingezet. Bij deepfakes gaat het dan zowel om de productie van deepfakes en het aanbieden van deepfaketechnologie als om mogelijke toetsen voordat deepfakes worden gegenereerd en/of verspreid.

5.1 Eerdere rapporten over horizontale privacy

Bij de vraag naar de handhaving van en het toezicht op bestaande en toekomstige rechtsregels kan allereerst worden gewezen op een aantal onderzoeksprojecten dat recentelijk voor het WODC is verricht, zoals een onderzoek naar het procesrecht in de data-gedreven samenleving en drie onderzoeken naar horizontale privacy. Ieder van die onderzoeken kwam grosso modo tot een vergelijkbare conclusie, namelijk dat alhoewel er materieelrechtelijk zeker op specifieke punten juridische aanpassingen mogelijk zijn, de grootste winst voor de horizontale privacy is te halen op het gebied van de handhaving en naleving van de regels.

- Het rapport over gezichtsherkenningstechnologieën in horizontale verhoudingen concludeert bijvoorbeeld dat de AVG en de UAVG hier zeer strakke grenzen aan stellen omdat hierbij gebruik wordt gemaakt van biometrische gegevens; omdat 'toestemming voor de verwerking van biometrische gegevens niet snel als uitzonderingsgrond zal kunnen dienen voor gezichtsherkenningstechnologieën [zal] het gebruik van dergelijke technologieën zelden legitiem [] zijn onder de AVG.'[372]
- Ten aanzien van het gebruik van drones en spionageproducten gelden tal van mogelijke strafrechtelijke bepalingen, kan het gebruik daarvan leiden tot een onrechtmatige daad en gelden er aanvullende regels die maken dat drones in principe niet mogen vliegen over een bebouwde kom of natuurgebied. Heimelijke opnames zullen vrijwel nimmer legitiem zijn,

omdat personen volgens de AVG van tevoren op de hoogte moeten worden gebracht van het feit dat er gegevens over hen zullen worden verzameld. 'Uit het voor dit rapport uitgevoerde onderzoek blijkt dat het grootste euvel ten aanzien van drones en spionageproducten niet zozeer is gelegen in juridische lacunes en onduidelijkheden, die er overigens wel op beperkte schaal zijn [], maar in een gebrek aan naleving en handhaving van de bestaande regels.'[373]

- In het onderzoek naar mogelijke lessen ten aanzien van de regulering van horizontale privacy uit andere landen werd geconcludeerd 'dat de normering van het recht op privacy in horizontale verhoudingen en de daarbij behorende rechtsbescherming in Nederland niet wezenlijk afwijkt van de wijze waarop deze is vormgegeven in de door ons onderzochte landen.'[374] De grootste winst, zo concludeerde dat onderzoek dan ook, zit in de procesrechtelijke en handhavingskant.
- Tot slot kwam ook het onderzoek ten aanzien van de hervorming van het procesrecht in de data-gedreven samenleving tot een soortgelijke conclusie. 'Tot nu toe is met name aandacht besteed aan de bescherming van de materiële rechten van burgers en lag de nadruk doorgaans op principes van materiële rechtvaardigheid. [] Minstens even belangrijk is echter dat er ook voldoende aandacht is voor de toegang tot het recht en principes van procedurele rechtvaardigheid. Immers, burgers die wel rechten hebben maar die niet met succes kunnen afdwingen, staan alsnog met lege handen. Ook kunnen, als het rechtsstelsel wel incidentele knelpunten adresseert, maar niet de structurele en onderliggende oorzaken, systeemfouten blijven bestaan. Deze vraagstukken hebben tot nu toe weinig aandacht gekregen.'[375]

Dat deze vier onderzoeken vergelijkbare uitkomsten hebben betekent uiteraard niet dat er geen winst zou zijn te behalen met aanpassingen aan het materiële recht. Maar het betekent wel dat de noodzaak om dergelijke aanpassingen te doen mede dient te worden beoordeeld vanuit het kernprobleem van privacyschendingen in horizontale verhoudingen, namelijk het gebrek aan toezicht, handhaving en naleving. Dat geldt zeker ook voor deepfakes. Aangezien de technologie gratis te downloaden is, materiaal (beeld, geluid, tekst) om deepfakes mee te maken vrijelijk beschikbaar is op het internet, er slechts beperkte technieken zijn om deepfakes bij voorbaat te weren en gepubliceerde deepfakes met eenvoud kunnen worden gekopieerd en verder verspreid, is slechts een beperkt deel van het probleem de juridische normering van deepfakes en het grootste probleem juist ligt gelegen in de handhaving van de materieelrechtelijke regels. Bovendien zal de keuze voor eventuele materieelrechtelijke regels in het licht moeten worden bezien van de handhaving en naleving. Bepaalde materiele bepalingen zijn immers beter te handhaven dan andere en *ex ante* regels zijn over het algemeen beter te handhaven dan *ex post* bepalingen.

Bij de handhaafbaarheid van regels 'op het internet', zeker in horizontale verhoudingen, zijn twee algemene punten van belang: de keuze voor *ex ante*- of *ex post*-regulering en de vraag wie de primaire normadressant dient te zijn.[376]

5.2 Ex ante- of ex post-regulering

In het huidige recht (strafrecht, AVG, civiel recht) is vrijwel uitsluitend gekozen voor *ex post*-regelgeving. Wat lijkt te pleiten voor meer

ex ante-regulering is onder meer dat alhoewel burgers zich in het algemeen bewust zijn van privacyregels – zeker de extremere schendingen die in het strafrecht zijn vervat zijn over het algemeen zo intuïtief dat ze als bekend kunnen worden beschouwd – niet elke burger op de hoogte is van het feit dat bij het maken en publiceren op een website van een deepfake waarop ook anderen te zien zijn, die anderen daarvan op de hoogte moeten worden gesteld en dat de AVG van toepassing is. Daarbij komt dat een burger die een legitiem doel nastreeft met een deepfake alsnog (onbewust) een onrechtmatige handeling kan verrichten. Meer in het algemeen zorgt *ex post* regulering ervoor dat de verzameling, verwerking en publicatie van informatie pas kan worden beoordeeld nadat het feit is geschied. Het probleem ten aanzien van privacy in horizontale verhoudingen is uiteraard dat steeds meer alledaagse producten burgers in staat stellen data van anderen te verzamelen en te verspreiden en dat het genereren van deepfakes met een handomdraai kan geschieden. Alhoewel het gebruik en toepassing van deepfaketechnologie nu nog beperkt is, valt het te verwachten dat op termijn miljoenen mensen in Nederland en daarbuiten een app of programma hebben gedownload die het produceren van deepfakes mogelijk maakt. Het kan dan op termijn ondoenlijk worden om bij iedere foto, video- of audiofragment te verifiëren of het hier om deepfake gaat en zo ja, na te gaan of de content rechtmatig is.

Dat brengt een aantal zaken met zich. Ten eerste dat de aandacht en energie vrijwel uitsluitend zal gaan naar de handvol extremere schendingen (vaak relaterend aan de lichamelijke privacy), maar dat het overgrote deel van vervelende, maar niet acuut problematische deepfakes ongemoeid blijft (het OM treedt slechts in een beperkt aantal zaken op; de Autoriteit Persoonsgegevens is vrijwel inactief als het gaat om horizontale privacyschendingen en richt zich op overheden en grote ondernemingen).[377] Dit zorgt er ook voor dat er op termijn een normalisering van deze kleine schendingen zal plaatshebben. Ten tweede dat mocht er wel iets aan de privacyschending worden gedaan, dat het dan reeds te laat is. De schade heeft zich al voltrokken, sterker nog, een rechtszaak kan zorgen dat er nog meer aandacht komt voor de ongewenste opnames (het zogenoemde Barbra Streisandeffect).[378]

Wat *ex ante* regulering echter lastig maakt is dat deepfakes ook voor legitieme doeleinden kunnen worden ingezet. *Ex ante* verboden maken ook deze legitieme toepassingen onmogelijk; *ex ante* toetsen op rechtmatigheid van toepassingen is vrijwel ondoenlijk gegeven de hoeveelheid van opnames en zal daarbij ook foutmarges met zich brengen, zoals Facebook dat in toenemende mate beelden verbiedt die een legitiem doel lijken te dienen. Dergelijke *ex ante* toetsen roepen natuurlijk ook de vraag op: welke partij beoordeelt of een product of toepassing legitiem is en op basis van welke juridische of morele standaard? Wat een *ex ante* toets op deepfakes bemoeilijkt is dat ze ook kunnen worden vervaardigd met de aanvankelijke toestemming van de betrokkene (een stelletje dat een fake-porno filmpje maakt van henzelf), maar dat de latere verspreiding daarvan (jongen wil wraak nemen op ex-vriendin) niet met toestemming geschiedt. Het is voor een derde (de Autoriteit Persoonsgegevens, een internet intermediair, etc.) vaak niet eenvoudig vast te stellen of beelden en/of de deepfake met toestemming

zijn gemaakt en zo ja, voor welke doeleinden toestemming is verkregen. Dat vaststellen is vaak een tijdrovende procedure. Daarbij komt dat alhoewel *ex ante* regulering beter te handhaven is, dit nog steeds geen garanties biedt, gezien de territoriale grenzeloosheid van het internet.

5.3 Normadressant

Drie partijen spelen een mogelijke rol bij het toezicht op en de handhaving van de privacynormen in horizontale verhoudingen – de burger zelf, de staat en de tussenpartijen –, maar er zijn belemmeringen ten aanzien van ieder van hen om dit effectief en adequaat te doen.

Iedere burger heeft weliswaar allerhande proces- en klachtrechten, maar het is lang niet altijd duidelijk voor iemand dat zijn data worden of zijn verzameld of dat er een deepfake van hem is verspreid op het internet (bijvoorbeeld op een pornosite of Geenstijl.nl). Zelfs als hij dat wel weet of te weten komt, dan nog is niet altijd duidelijk wie verantwoordelijk kan worden gehouden of aansprakelijk kan worden gesteld voor de mogelijke privacyschending. Zelfs bij een deepfake pornofilmpje dat twee geliefden met wederzijdse instemming hebben gemaakt, maar dat later tegen de zin van de één op het internet is verspreid, hoeft de ander daar niet per se achter te zitten. Zowel diens computer als de devices van het slachtoffer zelf kunnen bijvoorbeeld zijn gehackt, zodat een derde met het filmpje aan de haal kan zijn gegaan. Om achter diens identiteit te komen is vaak de medewerking van internet intermediairs noodzakelijk, maar die willen niet altijd meewerken (zonder last van de rechter) vanwege de privacybelangen van degene die het materiaal heeft geplaatst. Dat betekent dat er vaak twee rechtszaken nodig zijn, één om de identiteit van de dader te achterhalen en een andere om de dader in rechte aan te spreken. Als het daarbij nog gaat om een verwijderverzoek ten aanzien van het platform kan soms een derde rechtszaak nodig zijn. Dit vergt tijd, geld en energie die burgers vaak ontberen.

Partijen zoals Google, Facebook en Apple hebben zeer diepe zakken, zodat vaak slechts welgestelden een dergelijke langdurige en complexe rechtsgang kunnen volhouden.[379] De kosten en moeite van het voeren van dergelijke rechtszaken brengt met zich dat hoe dan ook de meer alledaagse privacyschendingen doorgaans niet geadresseerd zullen worden. Daarbij is voor de burger de vraag of bijvoorbeeld het enkele feit dat een deepfake van hem is gemaakt, zonder dat daar nu echt compromitterende zaken op te zien zijn, wel als onrechtmatig hebben te gelden, omdat de schade zo beperkt is. Zelfs als de schade wel kan worden aangetoond dan is het probleem dat die doorgaans laag zal zijn (afgezien van bijvoorbeeld fake pornobeelden), zodat een lange en vaak kostbare rechtsgang leidt tot een eventuele schadevergoeding van een aantal honderd euro.[380] Dat loont voor de burger nauwelijks de moeite.

Voor intermediairs geldt dat sommigen reeds deels een vorm van *ex ante* toetsing hebben doorgevoerd, maar dat dit zowel zeer tijdrovend als kostbaar is. Daarbij komt dat zij door dergelijke toetsen deels op de stoel van de rechter gaan zitten. Het probleem bij conflicten in horizontale verhoudingen is dat het zelden evident is of er een privacyschending heeft plaatsgevonden. Een internet intermediair

kan doorgaans niet aan een opname zien of die een onrechtmatig karakter heeft of niet. Een deepfake pornofilmpje kan met wederzijdse toestemming zijn gemaakt en verspreid. Doorgaans weet alleen de burger die getroffen wordt zelf dat er een vermoedelijke onrechtmatigheid aan een opname of publicatie ten grondslag ligt. Daarnaast gelden er tal van lastige juridische vragen. Worden er persoonsgegevens verwerkt bij een deepfake op basis van beelden van meerdere personen? Is er een legitieme verwerkingsgrondslag voor de deepfake? Geldt er een uitzondering in het kader van de vrijheid van meningsuiting? Vaak staan er voor verschillende burgers verschillende belangen op het spel, zodat de keuze om het verzoek van de één te honoreren tegelijk de keuze behelst om de rechten of belangen van de ander te beperken. Daarbij komt uiteraard dat dergelijke bedrijven vaak leven van de verkoop van producten of diensten en dat juist de controversiële content veel views en dus reclame-inkomsten genereren, zodat er een neiging is om zaken eerder wel dan niet te tolereren. Tot slot geldt voor intermediairs dat zij vaak in het (verre) buitenland zijn gevestigd. Het is voor hen vrijwel ondoenlijk om per regime weer een ander normatief stelsel door te voeren en te handhaven, waarbij de lastigheid ook nog is dat de twee burgers zich in verschillende jurisdicties kunnen bevinden, zoals een ingezetene in de Verenigde Staten die gebruik maakt van zijn vrijheid van meningsuiting en een burger binnen de Europese Unie die gebruik maakt van zijn recht op gegevensbescherming.

Voor overheidsinstanties, zoals het Openbaar Ministerie en de Autoriteit Persoonsgegevens, gelden veel van de voorgenoemde problemen evenzeer. De tijd, moeite en middelen die het kost om allerhande opnames en afbeeldingen op authenticiteit en rechtmatigheid te controleren, de onduidelijkheid die er vaak heerst over toestemming of niet en de diverse complexe juridische vragen die spelen, de vraag of het loont om achter vrij kleine privacyschendingen in horizontale verhoudingen aan te gaan en het probleem dat een keuze voor de bescherming van het recht van de ene burger vaak consequenties heeft voor de rechten van een andere burger. Voor overheidshandhaving is daarbij nog de vraag of het streng toezien op allerhande kleine privacyinbreuken niet erger is dan de kwaal. Zelfs als kosten noch moeite zouden worden gespaard, dan nog is de vraag of het wenselijk is om aan zulke vormen van repressie te doen. Overheidsdiensten meer macht en middelen geven om alledaags gebruik van deepfaketechnologie en applicaties te controleren kan leiden tot een *Big Brother* overheid die burgers nauwgezet in hun alledaagse bezigheden controleert.

5.4 Digital Services Act

Van belang is hierbij de regels voor de aansprakelijkheid en verantwoordelijkheid van internet intermediairs te benoemen. Die staan al meer dan 20 jaar in de e-Commerce Richtlijn,[381] maar worden op termijn waarschijnlijk vervangen door de Digital Services Act (DSA). Hieronder volgt een korte bespreking van de belangrijkste bepalingen uit het DSA-voorstel.[382]

Artikel 14 bevat de verplichting om een zogenoemd notice-en-takedown regime te implementeren en gebruikers de mogelijkheid te geven om providers ervan op de hoogte te stellen dat hun platform wordt misbruikt voor het delen van onrechtmatige informatie. Artikel

17 en 18 gaan over de mogelijkheid tot interne en buitengerechtelijke klachtenafhandeling. Interessant is artikel 19, die ziet op het aanstellen van 'flaggers'. '1. Onlineplatforms nemen de nodige technische en organisatorische maatregelen om te verzekeren dat via de in artikel 14 genoemde mechanismen door betrouwbare flaggers ingediende berichten prioritair en onmiddellijk worden verwerkt en afgehandeld. 2. De status van betrouwbare flagger krachtens deze verordening wordt, na aanvraag door een entiteit, toegekend door de coördinator voor digitale diensten van de lidstaat waarin de aanvrager is gevestigd, wanneer de aanvrager heeft aangetoond dat hij aan elk van de volgende voorwaarden voldoet: (a) hij heeft specifieke expertise en bevoegdheid voor het opsporen, identificeren en melden van illegale inhoud; (b) hij vertegenwoordigt collectieve belangen en is onafhankelijk van enig onlineplatform; (c) hij voert zijn activiteiten uit met als doel berichten tijdig, zorgvuldig en objectief in te dienen.'

Artikel 20 geeft verdere verplichtingen om misbruik van internetdiensten tegen te gaan. Zo moeten onlineplatforms stoppen met het verlenen van hun dienst aan gebruikers die stelselmatig misbruik maken en illegale inhoud delen. Daarbij dienen zij onder meer rekening te houden met: (a) de absolute aantallen manifest illegale inhoud of manifest ongefundeerde berichten of klachten die het voorbije jaar zijn ingediend; (b) het relatieve aandeel hiervan in verhouding tot de totale hoeveelheid aangeboden informatie of ingediende berichten in het voorbije jaar; (c) de ernst van het misbruik en de gevolgen ervan; (d) de intentie van de afnemer, persoon, entiteit of klager. Artikel 21 stelt dat de platforms politie en justitie moeten inlichten bij het vermoeden van bepaalde strafbare feiten. Tot slot bevat artikel 25 en verder verdere verplichtingen voor zeer grote online platforms, onder meer om goede risicoanalyses van hun diensten te verrichten en daar passende maatregelen op te treffen.

Noch deze voorgestelde Act noch de nu geldende e-Commercerichtlijn kennen een specifieke bepaling voor deepfakes. Het algemene regime is erop gericht om internet intermediairs vrij te stellen van aansprakelijkheid en hen vrij te waren van verplichtingen tot monitoring en filtering. Op deze benadering is veel kritiek. Het idee was aanvankelijk dat internetproviders, net als bijvoorbeeld postbedrijven, niet verantwoordelijk konden worden gehouden voor de post die zij verstuurden en ook vanwege het briefgeheim geen verplichting kon worden opgelegd om poststukken stelselmatig op inhoud te controleren. Enerzijds is zelfs in de analoge wereld een tendens om postbedrijven wel degelijk, zij het marginale, verplichtingen op te leggen. Anderzijds is van belang dat internet intermediairs een veel grotere en actievere rol zijn gaan spelen in het indexeren en vindbaar maken van materiaal en in het bepalen wat voor een soort materiaal er op de sites wordt gedeeld en hoe. Toch is ervoor gekozen om deze benadering in grote lijnen te handhaven; de ruimte voor Nederland om op dit punt strengere regels aan te nemen dan op EU-niveau is beperkt. Wel geldt dat als internet intermediairs van een mogelijke rechtsschending op de hoogte zijn, zij hierop actie moeten ondernemen.

Los van de vraag of dit in algemene zin een wenselijke aanpak is, is duidelijk dat deze benadering in ieder geval vragen oproept ten aanzien van deepfakes. Internet intermediairs hebben geen algemene plicht tot monitoren en filtering van materiaal, terwijl zij juist de meest voor de hand liggende partij zouden zijn

om al het binnengekomen materiaal door een 'deepfakescanner' te halen, die in ieder geval de meest duidelijke gevallen van deepfakes eruit zal halen. Alhoewel veel grotere partijen ongetwijfeld hier op vrijwillige basis actie op zullen ondernemen biedt het desalniettemin ruimte aan kwaadwillende partijen om van de regels gebruik te maken. Het is nu aan burgers zelf en aan eventuele burgerrechtenorganisaties om te ontdekken of reeds gepubliceerd materiaal een deepfake betreft of niet en zo ja, of die vermoedelijk als onrechtmatig dient te worden gezien. Dit terwijl zij vaak niet eens op de hoogte zijn van het bestaan van een deepfake over hen, laat staan de middelen hebben om voortdurend het hele internet af te speuren.

5.5 Conclusie

Dit hoofdstuk heeft een kort overzicht gegeven van een van de belangrijkste knelpunten in de horizontale privacybescherming in het algemeen en dat betreft de handhaving van en het toezicht op de materiele rechtsregels. Omdat veel van de regels momenteel zien op het gebruik van technologieën en niet op het produceren of aanbieden van technologieën, omdat doorgaans pas wordt geverifieerd of materiaal aan de wettelijke regels voldoet nadat dat is vervaardigd en verspreid en omdat het nagaan of bepaald materiaal al dan niet rechtmatig is, enkele uitzonderingen daargelaten, bij de burger is belegd, is er een praktijk ontstaan waarin de geldende regels veelvuldig worden overtreden, zonder dat hierop wordt geacteerd. Dat is niet verwonderlijk: per dag worden er alleen in Nederland al duizenden zo niet miljoenen foto's, video's en audiofragmenten gemaakt en verspreid via het internet. Die achteraf toetsten op rechtmatigheid en daarop eventuelele juridische stappen nemen is ondoenlijk.

Voetnoten hoofdstuk 5

- 372 Keymolen, E., Noorman, M., van der Sloot, B., Cuijpers, C., Koops, B. J., & Zhao, B. (2020). Op het eerste gezicht. Universiteit van Tilburg-Tilburg Institute for Law, Technology, and Society (TILT), Projectnummer: 2992, p. 135.
- 373 Galič, M., Noorman, M., van der Sloot, B., Koops, B. J., Cuijpers, C., Gellert, R., ... & van Delden, T. (2020). Spioneren met hobbydrones en andere technologieën door burgers: een verkenning van de privacyrisico's en reguleringsmogelijkheden, WODC, Projectnummer: 3063.
- 374 Schermer, B. W. & Van der Sloot, B. (2020), Het recht op privacy in horizontale verhoudingen WODC
- 375 Van der Sloot, B., & van Schendel, S. (2019). De modernisering van het Nederlands procesrecht in het licht van big data: Procedurele waarborgen en een goede toegang tot het recht als randvoorwaarden voor een data-gedreven samenleving, WODC, Projectnummer: 2900, p. 5
- 376 Onderstaand is gebaseerd op: Van der Sloot, B. (2021). 'Horizontale Privacy - Een probleemanalyse', Privacy & Informatie. Van der Sloot, B. (2021). 'Horizontale Privacy - De burger', Privacy & Informatie .
- 377 <https://www.autoriteitpersoonsgegevens.nl/nl/over-de-autoriteit-persoonsgegevens/focus-ap-2020-2023>.
- 378 <https://www.bbc.com/news/uk-18458567>.
- 379 Rechtbank Amsterdam, ECLI:NL:RBAMS:2019:8415, 11-11-2019.
- 380 Zie: Van der Sloot, B., & van Schendel, S. (2019).
- 381 Richtlijn 2000/31/EG van het Europees Parlement en de Raad van 8 juni 2000 betreffende bepaalde juridische aspecten van de diensten van de informatiemaatschappij, met name de elektronische handel, in de interne markt.
- 382 <https://eur-lex.europa.eu/legal-content/NL/TXT/PDF/?uri=CELEX:52020PC0825&from=en>.

6. Blik over de grens

In dit hoofdstuk wordt een blik geworpen op ontwikkelingen en reguleringsinitiatieven in het buitenland. Dat gebeurt in vier delen. Eerst wordt een quickscan geboden van relevante wetgevingsinitiatieven in het buitenland (paragraaf 6.1), dan wordt een samenvatting gegeven van twee voor deze studie uitgevoerde landenreportages naar de regulering van deepfakes in China (paragraaf 6.2) en de Verenigde Staten (paragraaf 6.3) (de volledige reportages, in het Engels, zijn te vinden in paragraaf 8.1) en wordt er een samenvatting gegeven van de belangrijkste uitkomsten van de veertien voor deze studie uitgevoerde interviews, met name met buitenlandse experts (paragraaf 6.4) (de volledige interviewverslagen zijn te vinden in paragraaf 8.2). Tot slot volgt een korte conclusie (paragraaf 6.5).

6.1 Quickscan reguleringsinitiatieven buitenland

In deze paragraaf wordt een kort overzicht gegeven van een aantal reguleringsinitiatieven wereldwijd. Het heeft niet als doel een uitputtend beeld te geven, maar geeft een algemeen beeld van verschillende manieren om deepfakes of de gevolgen daarvan te adresseren. Achtereenvolgens worden kort in het oog springende initiatieven besproken uit Australië, Canada, Duitsland, Filipijnen, Frankrijk, India, Oekraïne, Singapore en het Verenigd Koninkrijk.

6.1.1 Australië

De *Australian code of practice on disinformation and misinformation* is een zelfreguleringsinstrument aangenomen onder meer door Twitter, Google, Facebook, Microsoft, Redbubble en TikTok. Volgens de nieuwe code moeten ondertekenaars processen ontwikkelen voor het identificeren, beoordelen en verwijderen van verkeerde informatie en desinformatie op hun platforms. De code vereist dat deelnemers desinformatie verwijderen, accounts opschorten, misleidende inhoud labelen, een proces hebben voor het beoordelen van beslissingen die zijn genomen rond desinformatie, etc. Bedrijven die de code hebben ondertekend, moeten jaarverslagen publiceren over hoe ze de doelstellingen van de code trachten te behalen, met de eerste rapporten over de effectiviteit ervan in mei 2021.[383]

> *"The Australian Code of Practice on Disinformation and Misinformation has been adopted by Twitter, Google, Facebook, Microsoft, Redbubble, TikTok, Adobe and Apple.*
>
> *All signatories commit to safeguards to protect Australians against harm from online disinformation and misinformation, and adopting a range of scalable measures that reduce its spread and visibility.*
>
> *Participating companies also commit to releasing an annual transparency report about their efforts under the code, which will help improve understanding of online misinformation and disinformation in Australia over time. The first set of transparency reports were published on May 22, 2021 and are available below.*
>
> *DIGI developed this code with assistance from the University of Technology Sydney's Centre for Media Transition, and First Draft, a global organisation that specialises in helping societies overcome false and misleading information. The final code has been informed by a robust public consultation process.*

The Code was developed in response to the Australian Government policy announced in December 2019, where the digital industry was asked to develop a voluntary code of practice on disinformation, drawing learnings from a similar code in the European Union."[384]

Verder kent Australië een initiatief, Office of eSafety in Australia, om slachtoffers van non-consensuele deepnudes en seksueel getinte deepfakes een hulplijn te bieden door onder andere informatie te bieden over cyberrechten en hoe ze melding kunnen maken van de gemaakte deepnude/deepfake.[385]

6.1.2 Canada

De Canadese overheid heeft vooralsnog geen specifieke regulering ten aanzien van deepfakes aangenomen, maar benadrukt, net zoals onder meer in de Verenigde Staten, met name de politieke deepfakes en de bedreiging daarvan voor de democratie.

"New technology has created an emerging threat called deep fakes, which are synthetic videos often indistinguishable from real footage. Foreign adversaries can use this new technology to try to discredit candidates, and influence voters by, for example, creating forged footage of a candidate delivering a controversial speech or showing the candidate in embarrassing situations. Improvements in artificial intelligence (AI) are likely to enable interference activity to become increasingly powerful, precise, and cost-effective. Evolving technology underpinned by AI, such as deep fakes, will almost certainly allow threat actors to become more agile and effective when creating false or misleading content intended to influence voters, and make foreign cyber interference activity more difficult to detect and mitigate."[386]

6.1.3 Duitsland

In Duitsland wordt gekeken naar wetgeving waardoor rechercheurs gespecialiseerd in onderzoek naar seksueel misbruik van kinderen middels AI-gegenereerde 'kinderporno' kunnen fabriceren, om te dienen als lokmateriaal.[387] Op deze manier kunnen undercoveragenten infiltreren in kringen waar seksueel misbruik van kinderen plaatsvindt.[388] Dit project wordt naar verluidt samen met Microsoft uitgevoerd, die ook helpt bij het ontdekken van kinderporno.[389]

Duitsland kent verder geen specifieke regelgeving op federaal niveau die uitsluitend betrekking heeft op deepfakes. Wel is Duitsland bezig met campagnes rondom bewustwording van en weerbaarheid tegen desinformatie. Een voorbeeld hiervan is het DORIAN-project, waarin onder andere middels een podcast, schoolkinderen bewust gemaakt worden van de kenmerken van desinformatie.[390] Dit project richt zich op het ontdekken en bestrijden van desinformatie op het internet. De onderzoekers bij dit project hebben met behulp van Machine Learning bijvoorbeeld eigenschappen van kwaadwillige bots geïdentificeerd en fake inhoud automatisch gecategoriseerd op verschillende factoren zoals schrijfstijl. Ook hebben zij verdere strategieën ontwikkeld om deepfakes bloot te kunnen leggen, zoals de strategie om nepnieuws te onderzoeken op populistische patronen.[391]

Verder heeft het Bundestag de Netzdurchsetzunggesetz (netwerkhandhavingswet) aangenomen, die gericht is op nepnieuws op social media. Aanbieders van sociale media die meer dan

honderd klachten over onwettige inhoud ontvangen, zijn verplicht om halfjaarlijks Duitstalige rapporten op te stellen over de behandeling van klachten over onwettige inhoud op hun platform. De overheid probeert middels deze wet fake news en desinformatie (en zo ook deepfakes) te tackelen.

6.1.4 Filipijnen

In de Filippijnen voert een senaatscommissie momenteel een onderzoek uit naar wetgeving over deepfakes in relatie tot cybercrime- en gegevensbeschermingsrecht, mede omdat 'the proliferation of fabricated audios and videos [has] far-reaching consequences (...) on the protection of identity, personal information and privacy'. Er is een resolutie op dit punt aangenomen waarin het gevaar van deepfakes wordt benadrukt en tot slot wordt gesteld:[392]

> *"To direct the appropriate Senate Committee to conduct an inquiry, in aid of legislation, on the proliferation of Artificial Intelligence-synthesized audiovisual materials, otherwise known as deepfakes, with the end in view of strengthening government mechanisms to implement cybercrime and data privacy laws, safeguarding the privacy, information and identity of all Filipinos and protecting the integrity of Philippine social, political, economic and financial institutions."*[393]

6.1.5 Frankrijk

Frankrijk heeft sinds 2018 een wet (nr. 2018-1202 of 22 December 2018) die zich richt op de strijd tegen manipulatie van informatie. Deze wet staat bekend als de anti-nepnieuwswet en heeft vooral als doel eerlijke informatie aan burgers tijdens verkiezingen te garanderen. Het wordt gebruikt om de verspreiding van valse informatie tijdens de verkiezingsperiode te stoppen, meer transparantie van websites te waarborgen en de le Conseil Supérieur de l'Audiovisuel (vrij vertaald: de Hoge Audiovisuele Raad) krijgt meer regelgevende bevoegdheden. Artikel 11 uit de wet bepaalt:

> *"I. - Les opérateurs de plateforme en ligne mentionnés au premier alinéa de l'article L. 163-1 du code électoral mettent en œuvre des mesures en vue de lutter contre la diffusion de fausses informations susceptibles de troubler l'ordre public ou d'altérer la sincérité d'un des scrutins mentionnés au premier alinéa de l'article 33-1-1 de la loi n° 86-1067 du 30 septembre 1986 relative à la liberté de communication.*
> *Ils mettent en place un dispositif facilement accessible et visible permettant à leurs utilisateurs de signaler de telles informations, notamment lorsque celles-ci sont issues de contenus promus pour le compte d'un tiers.*
> *Ils mettent également en œuvre des mesures complémentaires pouvant notamment porter sur:*
> *1° La transparence de leurs algorithmes ;*
> *2° La promotion des contenus issus d'entreprises et d'agences de presse et de services de communication audiovisuelle ;*
> *3° La lutte contre les comptes propageant massivement de fausses informations ;*
> *4° L'information des utilisateurs sur l'identité de la personne physique ou la raison sociale, le siège social et l'objet social des personnes morales leur versant des rémunérations en contrepartie de la promotion de contenus d'information se rattachant à un débat d'intérêt général ;*
> *5° L'information des utilisateurs sur la nature, l'origine et les modalités de diffusion des contenus ;*
> *6° L'éducation aux médias et à l'information.*
> *Ces mesures, ainsi que les moyens qu'ils y*

consacrent, sont rendus publics. Chaque opérateur adresse chaque année au Conseil supérieur de l'audiovisuel une déclaration dans laquelle sont précisées les modalités de mise en œuvre desdites mesures.
II. - Le présent article est applicable en Polynésie française, dans les îles Wallis et Futuna, en Nouvelle-Calédonie et dans les Terres australes et antarctiques françaises."[394]

6.1.6 India

India kent geen speciale wetgeving aangaande deepfakes. Wel is het versturen van obsceen materiaal in elektronische vorm strafbaar. Deepfakes waarin vervalste versies van de originele inhoud worden gebruikt, kan een overtreding van artikel 468 IPC (Indian Penal Code) opleveren omdat dit artikel gaat vervalsing verbiedt. Ook als een deepfake wordt ingezet om op politiek niveau bijvoorbeeld haat te zaaien, dan kan dat op grond van artikel 124 IPC worden vervolgd.[395] Op Kamervragen over deepfakes gaf de Minister van Electronics and Information Technology het volgende antwoord:

"Yes, Sir. Government is aware of "Deepfake Technology" which enables to allows alter/ create the image or video media of a person with any other person's image or video using artificial neural networks and deep learning/machine learning techniques. The following measures have been taken by government to address such challenges of misinformation and rumours spreading via online platforms:

(i) Ministry of Electronics & Information Technology (MeitY) and Ministry of Home Affairs (MHA) are in regular touch with various social media platforms to effectively address the issue of removal of objectionable content, under the provisions of IT Act 2000.
(ii) Based on our emphasis, Social media platforms have implemented initiatives to address the issue of fake news on their platform, such as limiting forwards and promoting fact checking.

(iii) MeitY through a program, namely, Information Security Education & Awareness (ISEA), has been creating awareness among users and advising them not to share rumors/fake news. A dedicated website for information security awareness (https://www.infosecawareness.in) provides relevant awareness material.

(iv) MHA has created a Twitter Handle "Cyber Dost" to spread awareness on cyber safety and cyber security. MHA has also published a Handbook for Adolescents/Students on Cyber Safety.

The following are the major objectives of the proposed National Programme on Artificial Intelligence; with regard to tackling fake news:
• To establish a sustainable ecosystem for Artificial Intelligence in the country";
• To facilitate AI innovation, evolution, widespread use and implementation on bringing its benefits to citizens by preparing standards and enabling reforms including requisite changes in policies and laws
• To foster research and development in Artificial Intelligence and related technologies
• To build the capabilities in Artificial Intelligence at National and State level
• To bring out the global best practices and successful use cases for Indian Context
• To promote responsible AI deployment for the masses and generate public awareness"[396]

6.1.7 Oekraïne

In Oekraïne is een wetsvoorstel aanhangig betreffende de Wet op desinformatie, die tot doel heeft de verspreiding van desinformatie tegen te gaan. Wel ligt dit wetsvoorstel onder vuur, omdat deze wet het mogelijk zou maken om journalisten het zwijgen op te leggen en het tot aanzienlijke overheidscontrole op de media kan leiden. Het wetsontwerp voorziet in de functie van de regering om nepnieuws te identificeren en degenen die het verspreiden te straffen. Alle media, zo ook online en sociale media, zijn verplicht om persoonlijke informatie over journalisten te publiceren, inclusief hun namen en e-mailadressen. Het moet dienen als een soort 'vertrouwensindex'. Zodoende komen platforms en messenger-diensten onder meer controle van de overheid te staan.[397]

6.1.8 Singapore

De Protection from Online Falsehoods and Manipulation Act (POFMA) streeft ernaar om valse/misleidende informatie te voorkomen en het gebruik van online platforms om zulke informatie te kunnen communiceren tegen te gaan. De wet bevat een aantal maatregelen om de effecten van zulke communicatie tegen te gaan en om misbruik van online-accounts en bots te voorkomen. De wet richt zich onder meer op mededelingen die:

Schadelijk zijn voor de veiligheid van Singapore;

Schadelijk zijn voor de volksgezondheid, de openbare veiligheid, de openbare rust of de openbare financiën;

Nadelig zijn voor de vriendschappelijke betrekkingen van Singapore met andere landen;

De uitslag van een presidentsverkiezing, algemene verkiezing, tussentijdse verkiezing of referendum beïnvloeden;

Aanzetten tot gevoelens van vijandschap, haat of kwade wil tussen verschillende groepen personen; of

Het vertrouwen van het publiek in de overheid verminderen.

Voor overtredingen van bepaalde artikelen in de wet staat een maximale gevangenisstraf van vijf jaar.[398]

6.1.9 Verenigd Koninkrijk

De Online Harms Bill zal op termijn nieuwe aanknopingspunten bevatten om mensen te helpen veilig toegang te krijgen tot het internet. Dit betekent onder meer dat alle bedrijven die binnen het bereik van deze Bill vallen, actie moeten ondernemen om illegale activiteiten aan te pakken waarbij de veiligheid van kinderen in gevaar worden gebracht. Bedrijven die zich hier niet aan houden, riskeren een boete tot 10% van de omzet of maximaal 18 miljoen pond. Hieronder vallen ook de grote bedrijven, zoals Facebook/Instagram en Twitter. Deze Bill richt zich met name op desinformatie, bijvoorbeeld de desinformatie die bepaalde groepen verspreiden over COVID-19.[399]

6.2 China

China heeft diverse wetten die deepfakes aan nadere regels onderwerpen. Deze staan uitgebreid beschreven in het landenrapport dat voor deze studie is geschreven door Bo Zhao. Deze studie is te vinden in bijlage 8.1. Hieronder volgen kort een aantal van de belangrijkste bevindingen.

Allereerst legt de nieuwe Network Audio-video Information Services Regulation aanbieders van audio- en videodiensten expliciet specifieke verplichtingen op om bij desinformatie

gebruikte deepfaketechnologie aan te pakken door dergelijke informatie of inhoud duidelijk te identificeren en te markeren, en verbiedt het expliciet het gebruik van deepfakes bij de productie en verspreiding van nieuwsberichten.

Anders dan de EU-wetgeving inzake gegevensbescherming, beschermt art. 994 van het Chinese burgerlijk wetboek de persoonlijkheidsrechten van de overledene, met inbegrip van hun namen, gelijkenis, reputatie, eer, privacy en dergelijke, en hebben de echtgenoot, de kinderen, de ouders en de familieleden van de overledene het recht om te verzoeken de burgerlijke aansprakelijkheid te dragen overeenkomstig de wet. Dit betekent dat in geval van misbruik of oneigenlijk gebruik van de beelden en stemmen van overledenen in deepfakes, hun familieleden en nakomelingen rechtsmiddelen kunnen aanwenden in een rechtsgeding.

In het kader van de aansprakelijkheid op grond van onrechtmatige daad heeft het nieuwe burgerlijk wetboek gedetailleerde clausules vastgesteld voor inbreuken tegen Chinese burgers door netwerkgebruikers en dienstverleners. Art. 1194 schrijft voor dat internetgebruikers en dienstverleners civielrechtelijk aansprakelijk zijn wanneer zij door middel van het internet inbreuk maken op de burgerrechten van anderen. Art. 1195 verduidelijkt dat wanneer netwerkgebruikers het internet gebruiken om tegen anderen een dergelijke inbreuk te plegen, laatstgenoemden het recht hebben de aanbieder van netwerkdiensten in kennis te stellen van de nodige maatregelen, waaronder het wissen, blokkeren en deblokkeren van de verbinding. Een dergelijke kennisgeving bevat het primaire bewijs van de inbreuk en de werkelijke persoonlijke gegevens van de overtreder. Na ontvangst van de kennisgeving geven de netwerkdienstverleners de kennisgeving door aan de netwerkgebruiker en nemen zij tijdig de nodige maatregelen naar gelang van de inbreuken; zo niet, dan neemt de dienstverlener de gezamenlijke verantwoordelijkheid op zich.

De belangrijkste regels aangaande deepfakes staan in de Verordening betreffende netwerkaudio-video-informatiediensten.

Ten eerste pakt de Verordening deepfakes rechtstreeks aan. Art. 11, lid 2, verbiedt uitdrukkelijk 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. De Verordening schrijft voor dat reproductie van of verwijzing naar audio-/video-nieuws of -informatie alleen mag plaatsvinden op basis van audiovisuele nieuwsinformatie die is gemaakt door de instellingen die daarvoor toestemming hebben gekregen in de desbetreffende wetten.

Ten tweede legt de Verordening belangrijke regelgevende verplichtingen op aan aanbieders van audio-videodiensten. Aanbieders van netwerk-audio/videodiensten krijgen regelgevende taken opgelegd met betrekking tot het monitoren, filteren en reguleren van dergelijke audio-video-informatie/inhoud. Aanbieders van netwerk-audio/videodiensten worden gedefinieerd als entiteiten die audio/videowerken produceren, distribueren en verspreiden via netwerkplatforms zoals het internet en apps. Art. 6 eist dat aanbieders van audio/videodiensten eerst een exploitatiecertificaat of -kwalificatie moeten verwerven om dergelijke diensten te mogen aanbieden (als voorwaarde om met

hun bedrijf te mogen beginnen). Dit legt een hogere drempel op om de markt te betreden, en regelgevende instanties kunnen dienstverleners diskwalificeren wanneer zij niet voldoen aan de opgelegde verplichtingen.

Art. 10 vereist de uitvoering van een veiligheids- en risicobeoordeling (volgens de desbetreffende nationale wetgeving) door aanbieders van online audio/videodiensten die gebruikmaken van deep learning- en virtual reality-technologieën en vereist dat dergelijke content duidelijk gemarkeerd moet worden.

Art. 12 stelt dat diensten audio-/video-informatie moeten controleren, onder meer door: a) de installatie van controle- en identificatietechnologieën, b) het stopzetten van de doorgifte van illegale of onwettige informatie, het nemen van verdere maatregelen zoals het wissen en voorkomen van verspreiding, het bijhouden van desbetreffende gegevensbestanden, enz. en c) het melden van dergelijke inhoud aan regelgevende instanties op het gebied van netwerken, cultuur en omroep. Zodra aanbieders van netwerkdiensten informatie of inhoud aantreffen die illegaal is, moeten zij de verwerking daarvan stopzetten en de verdere verspreiding blokkeren.

Art. 13 verzoekt dienstverleners om, zodra zij vaststellen dat gebruikers van audio/videodiensten desinformatie (geruchten) uiten, verspreiden of publiceren door gebruik te maken van valse video/audioproducten op basis van deep learning en virtuele reality, onmiddellijk maatregelen te nemen om de situatie op te helderen en verslag uit te brengen aan de betrokken regelgevingsinstanties met het oog op archivering.

Ten derde schrijft de verordening ook concrete maatregelen voor die aanbieders van audio-/video-informatiediensten moeten volgen om deepfakes en andere misinformatie op basis van deep learning- en virtual reality-technologieën te bestrijden. In het bijzonder vraagt art. 8 dat de gebruikers moeten worden geïdentificeerd en vraagt Art. 14 voorts om duidelijke contractuele clausules ter verduidelijking van de plichten en rechten van zowel de aanbieders van audio-video-informatiediensten als de gebruikers van hun diensten, met name de verplichting van de gebruikers om zich te houden aan de Chinese wet- en regelgeving; ook hebben de aanbieders van informatiediensten de bevoegdheid om de nodige maatregelen te nemen, bijvoorbeeld om diensten te beperken, op te schorten of stop te zetten, een waarschuwing te sturen, gegevens te archiveren of verslag uit te brengen aan de bevoegde overheidsinstanties.

Om de tenuitvoerlegging van de bovenstaande regels te waarborgen, schrijft artikel 18 sanctiemaatregelen voor overeenkomstig andere wetten en regels en wordt verduidelijkt dat dit administratieve sancties en strafrechtelijke verantwoordelijkheden kan inhouden.

6.3 Verenigde Staten

In de Verenigde Staten zijn er, voornamelijk op statelijk niveau, diverse wetten die deepfakes aan nadere regels onderwerpen. Deze staan uitgebreid beschreven in het landenrapport dat voor deze studie is geschreven door Andrew Roberts. Deze studie is te vinden in bijlage 8.1 Hieronder volgen kort een aantal van de belangrijkste bevindingen.

Uiteraard is binnen de Verenigde Staten, het Eerste Amendement, waarin is vervat de vrijheid van meningsuiting van zeer groot belang. De VS kennen een zeer brede definitie van free speech en leggen slechts beperkte grenzen op. Toch zijn die er wel, bijvoorbeeld ten aanzien van onjuiste uitingen. Deze uitingen zijn in het politieke debat wel aan minimale regels gebonden, omdat regels anders de overheid in staat zou stellen om als scheidsrechters van waarheid en leugens op te treden, en de staat - via zijn organen voor strafvervolging - de macht zou krijgen over diegenen wier opvattingen niet stroken met de heersende ideologie. Maar false "meningsuiting" (false speech) die juridisch aantoonbare schade veroorzaakt, zal geen bescherming genieten, en wetgeving die strafrechtelijke of civielrechtelijke aansprakelijkheid oplegt met betrekking tot dergelijke onjuiste voorstellingen zal niet in strijd met het Eerste Amendement worden geacht te zijn. Bijgevolg kan de wetgever een vordering mogelijk maken wegens aantasting van de goede naam door lasterlijke uitlatingen, of uitlatingen bestraffen die bedoeld zijn om aan te zetten tot geweld zonder het Eerste Amendement te schenden.

De bescherming van het Eerste Amendement strekt zich niet uit tot bedrieglijke voorstellingen die de integriteit van de openbare instellingen en het vertrouwen van het publiek in die instellingen zouden kunnen ondermijnen, noch tot valse verklaringen die bedoeld zijn om anderen te bedriegen met eigendommen of andere zaken van waarde, zoals kansen op een baan. Wetgeving die voorziet in strafrechtelijke aansprakelijkheid voor het zich voordoen als overheidsambtenaar of het afleggen van valse getuigenissen in gerechtelijke procedures zal niet in strijd zijn met het Eerste Amendement. Om onder de uitzondering op de algemene bescherming te vallen, moet echter worden aangetoond dat er een direct causaal verband bestaat tussen de opgelegde beperking en de schade die moet worden voorkomen. Dit zal bij deepfakes niet altijd eenvoudig zijn.

Er is op een aantal punten nadere regelgeving aangenomen. Daarbij gaat het met name om regels die deepfake porno moeten tegengaan en regels die deepfake nepinformatie moeten tegengaan tijdens verkiezingen. Ook zijn er regels over de commerciële exploitatie van portretrechten van overleden personen.

Drie staten hebben 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 wettelijke privaatrechtelijke grond tot het instellen van een vordering biedt. Sectie 1708. 86 van het Californische burgerlijk wetboek bepaalt dat een persoon die als gevolg van "digitalisering" een voorstelling lijkt te geven die hij niet daadwerkelijk heeft gegeven, of die een "gewijzigde afbeelding" geeft, een vordering kan instellen tegen (i) degenen die expliciet seksueel materiaal creëren en opzettelijk openbaar maken, en die weten of redelijkerwijs hadden moeten weten dat de persoon die in het materiaal wordt afgebeeld, niet heeft ingestemd met de creatie of openbaarmaking ervan; en (ii) eenieder die expliciet seksueel materiaal opzettelijk openbaar maakt, wetende dat het is gecreëerd zonder de toestemming van de persoon die erin wordt afgebeeld.

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. Het ongeoorloofd gebruik van een "digitale kopie" is een grond voor het instellen van een rechtsvordering. De wet voorziet in rechtsmiddelen in geval van ongeoorloofd gebruik van de beeltenis van een overleden uitvoerende kunstenaar in een artistiek werk dat met behulp van deepfaketechnologie is vervaardigd. Dit wordt gedaan door het mogelijk te maken een schadevergoedingsactie in te stellen tegen eenieder die (i) "zonder voorafgaande toestemming een digitale replica van een overleden uitvoerende kunstenaar gebruikt in een gescript audiovisueel werk als fictief personage of voor de live uitvoering van een muziekwerk", (ii) wanneer de deepfake "het publiek de indruk kan geven dat het werk geautoriseerd is". In de bepalingen wordt uitgelegd dat het recht op publiciteit een eigendomsrecht is dat "vrij overdraagbaar is, geheel of gedeeltelijk, bij overeenkomst, bij schenking, of bij wege van een trust, of bij wege van enig ander testamentair instrument". Elke persoon die de rechten bezit op het materiaal dat is gebruikt om de deepfake te creëren, kan schadevergoeding vorderen voor de schade die is veroorzaakt door het ongeoorloofde gebruik ervan. Deze schadevergoeding kan alle winst omvatten die aan het ongeoorloofde gebruik kan worden toegeschreven, alsmede een schadevergoeding als straf. De vordering verjaart veertig jaar na het overlijden van de overledene.

Er zijn een aantal omstandigheden waarin ongeoorloofd gebruik geen aanleiding geeft tot aansprakelijkheid. De eerste is wanneer de persoon die de afbeelding gebruikt in de digitale replica "een opvallende disclaimer in de aftiteling van het gescripte audiovisuele werk, en elke daarmee verband houdende reclame waarin de digitale replica verschijnt, met de vermelding dat het gebruik van de digitale replica niet is toegestaan". Er zijn ook uitzonderingen voor het gebruik van digitale replica's in parodiërende werken, satire, kritiek en commentaar, en; documentaires, historische en biografische werken. Andere omstandigheden waarin degenen die een digitale replica gebruiken niet aansprakelijk zijn, zijn onder meer nieuwsverslaggeving, openbare aangelegenheden en sportprogramma's.

Er is een aantal staten dat regels heeft gesteld ten aanzien van het verspreiden van nepinformatie ten tijde van verkiezingen. Zo stelt een Texaanse wet het maken en publiceren van materiaal strafbaar dat is bedoeld om de uitslag van een verkiezing te beïnvloeden. De wet gebruikt daar de term "deepfake video": "deepfake video" betekent een video (...) die een echte persoon lijkt weer te geven die een handeling verricht die niet in werkelijkheid heeft plaatsgevonden". Ook de staat Californië heeft wetgeving uitgevaardigd om het probleem aan te pakken van bedrieglijke audiovisuele opnamen die worden gemaakt en gebruikt met de bedoeling de verkiezingen te beïnvloeden. De aanpak van deze staat verschilt echter van die van de Texaanse wetgever. In plaats van te trachten dergelijk gedrag te ontmoedigen door de dreiging van strafrechtelijke sancties, heeft zij een wettelijke privaatrechtelijke grond voor een vordering ingesteld die kan worden ingeroepen door kandidaten bij verkiezingen die het doelwit zijn geweest van gemanipuleerde audio- of video-opnamen.

De Texaanse wet is interessant omdat het een definitie geeft van een deepfakevideo. Een dergelijke definitie is ook te vinden in een gestrande wet op landelijk niveau. De wet bevatte voorstellen voor een reeks strafbare feiten in verband met bepaalde vormen van gebruik van een "geavanceerde technologische valse personificatieregister". Het definieerde een "geavanceerd technologisch fictief personificatieregister" als:

> *'Any deep fake which –*
>
> *(A) a reasonable person, having considered the audio or visual qualities of the record and the nature of the distribution channel in which the record appear, would believe accurately exhibits –*
>
> *(i) any material activity of a living person which such living person did not in fact undertake; or*
>
> *(ii) any material activity of a deceased person which such deceased person did not in fact undertake, and the exhibition of which is substantially likely to further a criminal act of result in improper inference in an official proceeding, policy debate, or election; and*
>
> *(B) was produced without the consent of such living person, or in the case of a deceased person, such person or heirs thereof.*

Het vervolgt dat te stellen dat *'material activity'* is:

> *'any falsified speech, conduct or depiction which causes, or a reasonable person would recognise has a tendency to cause perceptible individual or societal harm, including misrepresentation, reputational damage, embarrassment, harassment, financial losses, the incitement of violence, the alteration of a public policy debate of election, or the furtherance of any unlawful act.'*

'Deepfake' wordt gedefinieerd als:

> *'any video recording, motion-picture film, sound recording electronic image, or photograph, or any technological representation of speech or conduct substantially derived thereof –*
>
> *(A) which appears to authentically depict any speech or conduct of a person who did not in fact engage in such speech or conduct; and*
>
> *(B) the production of which was substantially dependent upon technical means, rather than the ability of another person to physically or verbally impersonate such person.*

6.4 Interviews

Er zijn vijftien interviews afgenomen. Elf interviews zijn gehouden met internationale experts. Daarbij is gekozen voor een diversiteit aan professionele en wetenschappelijke achtergronden – geïnterviewden hadden onder meer een technische achtgrond, een wetenschappelijke titel in communicatiewetenschap, waren jurist, hadden een economische achtergrond of specialiseerden zich in vrouwenstudies. Bij de selectie van kandidaten is ook rekening gehouden met de territoriale spreiding, om zo een beeld te krijgen van mogelijk verschillen tussen de verschillende continenten en regio's, zowel qua ervaringen met deepfakes, de benadering van dit fenomeen als ten aanzien van de daar vigerende wetgeving. Ook zijn vier interviews gehouden met Nederlandse experts op het gebied van het

procesrecht (aansluitend op hoofdstuk 4): twee werden gehouden met personen met kennis van het burgerlijk procesrecht, twee met kennis van het strafprocesrecht. De volledige verslagen van de interviews zijn te vinden in paragraaf 10.2, hieronder volgt een samenvatting op een viertal punten: de techniek, de toepassing en de gevolgen daarvan, de specifieke inzichten ten aanzien van het mogelijk gebruik van deepfakes in de rechtszaal en de reguleringsmogelijkheden ten aanzien van deepfakes. De gehouden interviews zijn met de volgende personen:

- 1
  - **Judit Altena-Davisen**
  - Strafrechtjurist; Nederland
- 2
  - **Margreet Ashmann**
  - Civilist; Nederland
- 3
  - **Ruth de Bock**
  - Civilist; Nederland
- 4
  - **Jacquelyn Burkell & Chandell Gosse**
  - Information & Media Studies; Canada
- 5
  - **Manon den Dunnen**
  - Strategisch specialist digitaal, Nationale Politie; Nederland
- 6
  - **Serena Iacobucci**
  - Behavioral Economics; Italië
- 7
  - **Tyrone Kirchengast**
  - Strafrechtjurist; Australië
- 8
  - **Andrei Kwok Onn Jui**
  - Management; Maleisië
- 9
  - **Hao Li**
  - Computer Scientist; Verenigde Staten
- 10
  - **Sophie Maddocks**
  - Media & Communication; Verenigde Staten
- 11
  - **Emma Perot**
  - Commerical law; Trinidad & Tobago
- 12
  - **Lonneke Stevens**
  - Strafrechtjurist; Nederland
- 13
  - **Aya Yaldin**
  - Politics & Communication; Israel
- 14
  - **Mika Westerlund**
  - Technology Innovation Management; Canada
- 15
  - **Christopher Whyte**
  - Political Science; Verenigde Staten

6.4.1 Technische middelen en tegen middelen

Alle geïnterviewden verwachten dat deepfaketechnologie zich in de komende jaren snel zal ontwikkelen. Er wordt vaak gewezen op het feit dat de techniek pas een handvol jaren oud is en dat sindsdien de capaciteiten en toepassingsmogelijkheden met rassenschreden vooruit zijn gegaan. Een deepfake is steeds makkelijker te produceren tegen steeds lagere kosten. Belangrijk is wel dat er geen wezenlijk nieuwe toepassingsgebieden of kwalitatieve veranderingen worden verwacht, maar “slechts” kwantitatieve. Deepfakes zullen nog eenvoudiger te maken zijn, de techniek zal nog goedkoper worden in aanschaf en gebruik, de techniek zal nog echtere content produceren, daarbij zullen nog minder artefacten, dat wil zeggen sporen van fabricage, worden achtergelaten en de techniek zal in nog meer handen komen, zo niet vrijwel de hele burgerbevolking.

Deze ontwikkeling spiegelt met de inzichten ten aanzien van deepfake-detectietechnologie. Interessant is dat hier een scheiding valt te zien tussen de geïnterviewden met een juridische achtergrond en geïnterviewden met een technische achtergrond.

De eerste groep verwacht veel van technische detectiemiddelen, wijst er op dat er altijd een kat-en-muisspel is tussen technologie en tegen-technologie, geven aan dat zij de verwachting hebben dat er flink geïnvesteerd zal worden in tegen-technologieën en dat het vermoedelijk net zoals met antivirussoftware zal gaan: die kan niet alle virussen er uit pikken, maar is een standaard geworden voor iedere computer en digitaal apparaat en is in staat veruit de meeste apparaten tegen infectie te beschermen.

De tweede groep is veel sceptischer, om niet te zeggen pessimistisch. Zij wijzen erop dat detectietechnologieën nu al slechts 65% van de deepfakes eruit kunnen halen (ondanks de jubelende krachtenberichten die soms verschijnen, die vaak een misleidende voorstelling van zaken geven, bijvoorbeeld omdat de deepfake-detectietechniek is ingezet in een afgesloten testsetting met een beperkte set aan deepfakes). Detectiemiddelen zullen nimmer boven dit percentage komen, stellen zij, eerder zal dit percentage minder worden. Bovendien wijzen zij erop dat het meest waarschijnlijke scenario is dat dergelijke technieken uiteindelijk 'waarheidspercentages' of 'betrouwbaarheidspercentages zullen geven: de kans dat deze video authentiek/niet gemanipuleerd is, is 73%, en geen uitsluitsel zal geven. De beste strategie om deepfakes te ontdekken, stellen zij, is niet door middel van tegen-technologie, maar door middel van een menselijke inschatting naar de waarheid op basis van contextuele informatie: is dit iets wat deze persoon normaal gesproken zou zeggen? Sluit dit aan bij andere informatie omtrent dit onderwerp of deze persoon? Zijn er andere bronnen die hetzelfde berichten, etc. Toch is de nieuwste ontwikkeling waarop zij wijzen precies hierop gericht. AI wordt ingezet om hele fake omgevingen in te richten: niet alleen een deepfakevideo, maar ook fake nieuwswebsites die daar een reportage over maken, fake Twitter-accounts die met elkaar in discussie gaan over het filmpje, fake Insta-accounts waar memes van het filmpje worden gegeneerd, Wikipagina's die automatisch worden aangepast of aangemaakt over dit onderwerp, fake-journaalitems die over de ophef berichten. Zo ontstaat een hele omgeving van nepinformatie, waardoor het zowel voor mens als algoritme steeds lastiger wordt om echt van nep te onderscheiden.

Daarbij wijzen experts erop dat steeds meer van de digitale content in meer of mindere mate gemanipuleerd is. De verwachting is dat over zes jaar meer dan 90% van alle digitale content gemanipuleerd is. Vaak gaat het daarbij om relatief kleine aanpassingen. Zo zit er een bepaalde geluidscompressie op de meeste opnameapparatuur en video-belsystemen, waardoor niet alle geluiden worden opgevangen en de opgevangen geluiden vervormd worden doorgegeven. Ook zit op steeds meer smarttelefoons en andere apparaten standaard een techniek ingebakken die de huid van de afgebeeldene egaler maakt. Dat lijkt relatief onschuldig, maar zelfs dit soort kleinere manipulaties kunnen van groot belang blijken, bijvoorbeeld bij de identificatie van een verdachte door een getuige of bij een medische behandeling met een video-consult. Daarnaast

zijn er natuurlijk ook de deepfakes, die een wezenlijk andere voorstelling van zaken geven.

6.4.2 Toepassing en consequenties

Over de definitie van deepfakes bestaat geen eenduidigheid onder de geïnterviewden. Sommigen wensen het fenomeen te beperken tot video- en beeldmanipulatie met behulp van GAN-technologie, anderen zien het breder: het gaat uiteindelijk over de manipulatie van materiaal door moderne technologie. Bij materiaal kan het gaan om beeld en geluid, maar ook om tekst en bijvoorbeeld satellietsignalen. Het gaat dan om eigenlijk iedere vorm van manipulatie door middel van technische hulpmiddelen, waarbij juist ook de fake-tekstberichten volgens experts een grote nieuwe dreiging zullen vormen. Alhoewel fake-beelden makkelijker 'binnenkomen' bij mensen, hebben mensen bij geschreven bronnen sneller de neiging om de echtheid daarvan voor waar aan te nemen. Ook zijn deze fake-teksten nog makkelijker te genereren. Wel is van belang dat waar het bij beelden in traditionele zin doorgaans gaat om bestaande beelden die worden gemanipuleerd, het bij teksten vaak gaat om volledig nieuw gegenereerde woorden.

Qua toepassingen van deepfakes wordt er gewezen op het gevaar, dat meer in algemeen speelt bij technologie, om in utopieën of in dystopieën te denken. Juist omdat beleidsmakers, journalisten en burgers vaak weinig snappen van de techniek worden de toepassingsmogelijkheden van moderne middelen vaak in extremen uitgedrukt. Utopisten menen dat de techniek de wereld efficiënter, veiliger en leuker gaat maken, pessimisten vrezen voor een 'killer-robot' en technieken die onze waarden en rechtsstaat fundamenteel ondermijnen. Ook bij deepfakes wordt er veel ingezet op de mogelijk kwaadaardige kanten van deepfakes, die het goed doen in de media. Het stoort velen dat er zo veel aandacht wordt besteed aan mogelijk politieke toepassingen, bijvoorbeeld Rusland die middels deepfakes de democratische verkiezingen in een land proberen te beïnvloeden, terwijl dit nochtans een zeer klein toepassingsgebied van de deepfaketechnologie betreft. Sommigen stellen echter dat het wel van belang is om op dit punt goed de vinger aan de pols te houden, omdat de mogelijke gevolgen significant kunnen zijn.

Qua toepassingen werd er door velen op gewezen dat zo'n 95% van de deepfakes porno betreft en wel non-consensual porn, dus zonder toestemming van de afgebeeldene(n). Als 95% van een bepaalde techniek voor een toepassing wordt gebruikt en die toepassing zeer schadelijk en negatief is, dan moet die toepassing eigenlijk ook in de definitie van deepfake worden opgenomen. Een auto kan je ook als museumstuk of verzamelobject gebruiken, maar veruit de grootste toepassingscasus van auto's is het van A naar B rijden. Deze toepassing wordt dan ook in de definitie van auto's opgenomen, namelijk dat het primair een vervoermiddel is. Net zozeer wordt deepfaketechnologie primair ingezet voor het vervaardigen van porno. Als definitie stelde mensen dan ook voor 'deepfake is a technology....., primarily used for the creation of fake, non-consensual porn'.

Interessant is dat anderen juist een diametraal tegenovergestelde positie innamen. Techniek is neutraal, zo zeiden zij, het is maar wat je er mee doet. Daaruit volgt dat techniek ook niet dient te worden verboden, omdat techniek niet intrinsiek goed of slecht is en je daarmee ook de goede use cases zou verbieden. Het gaat er in deze

visie om de slechte toepassingen door mensen in te kaderen. De eerste groep wijst er juist op dat techniek nooit neutraal is. Er gaan bepaalde use cases uit van het ontwerp van een bepaalde techniek. Een keukenmesje kan worden gebruikt om een moordaanslag te plegen, maar dat is bepaald niet eenvoudig, terwijl het uitermate geschikt is voor het snijden van komkommer. Een vuurwapen is voor dat laatste weer ongeschikt, terwijl het primair is ontworpen voor het doden of verwonden van levende wezens. Het komt neer op het wapendebat in de Verenigde Staten. Aan de ene kant stelt men 'Guns don't kill people, people kill people'; aan de andere kant stelt men: 'but the gun helps'.

Waar alle geïnterviewden het over eens zijn is dat de primaire 'winst' of positieve toepassing van de deepfaketechnologie gelegen is in de meer professionele toepassingen, zoals binnen de film- en entertainmentindustrie, het gebruik voor medische toepassingen, het adresseren van minderheden of de wereldbevolking in hun eigen taal of het beeldbellen met zakenpartners in overzeese gebieden, waarbij niet alleen het gesprokene on the spot wordt vertaald en als audiofragment aan de andere kant belandt, maar ook de lippen van de spreker zo worden aangepast dat het lijkt alsof hij de woorden in de vreemde taal ook echt heeft gebezigd. Het genereren van deepfakes kan ook worden gebruikt om personen te anonimiseren, ofwel binnen de medische context ofwel binnen de strafrecht context, bijvoorbeeld om getuigen die anoniem willen blijven te beschermen. Ook kunnen sekswerkers zelf deepfakes gebruiken in hun werkzaamheden, al wordt erop gewezen dat het met name hun 'content', hun films, zijn die worden gebruikt voor het vervaardigen van deepfake-porno, waarbij hun lichaam en handelingen geportretteerd blijven, maar hun hoofd wordt vervangen door die van een ander. Een controversiële toepassing is het gebruik van deepfake kinderporno, van niet bestaande kinderen, voor de behandeling van pedofielen of de opsporing van pedoseksuelen. Voor consumententoepassingen wordt er met name gewezen op vermaak en satire.

Qua gevaren wordt er naast de negatieve aspecten van fakeporno en het destabiliseren van democratieën met name gewezen op het verspreiden van misinformatie, voor welk doeleinde dan ook. Ook kunnen deepfakes worden ingezet voor frauduleuze handelingen en misleiding. Tot slot zijn er voorbeelden van gewelddadigheden ten aanzien van personen die iets naars lijken te doen op een deepfakevideo, terwijl dat niet daadwerkelijk is gebeurd, zoals een vrouw die een dier lijkt te mishandelen. Dit heeft de vrouw in kwestie in fysiek gevaar gebracht. Bij porno werd er nog gewezen op het feit dat de echtheid van een filmpje vaak niet uitmaakt. Een persoon die zich wil verlekkeren aan een pornofilm met zijn favoriete celebrity weet dat het fake is. Als er een fake porno-filmpje wordt gemaakt van een pubermeisje, dan kan zij en haar hele klas weten dat iets fake is en toch kan het de sociale dynamiek in de klas veranderen en het zelfbeeld van het meisje aantasten. Het zien van jezelf in bepaalde posities, ook al weet je dat het nep is, doet iets met je perceptie en zelfwaarde. Ook voor misinformatie geldt dit principe deels. Mensen willen vaak filmpjes en berichten geloven die in hun politieke overtuiging of wereldbeeld passen. Als het bericht achteraf nep blijkt te zijn hechten zij daar weinig waarde aan: het bericht mocht dan wel nep zijn, maar de onderliggende waarheid of strekking klopt nog steeds.

Het is daarbij wel van belang dat verschillende respondenten, ook zij die met name overtuigd waren van het gevaar van deze technologie, er op wezen dat het 'werkelijke' probleem breder en maatschappelijker van aard was. Deepfake pornofilmpjes zijn in feite een uitvloeisel van het disrespect voor vrouwen en het objectiveren van het vrouwenlichaam dat 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 uitten in tal van hacks en spionageactiviteiten. Zelfs in een wereld zonder deepfakes zouden deze tendensen zich waarschijnlijk doorzetten.

Het nieuwe en potentieel gevaarlijke van deepfakes zat volgens de respondenten in een tweetal zaken, een kwalitatief en de ander kwantitatief. Enerzijds lijken deepfakes zo echt dat zij sneller voor waar zullen worden gehouden. Mensen hebben een zogenoemde 'truth-bias', ze nemen aan dat iets waar is, tenzij er tegenargumenten zijn. Dat geldt zeker voor videobeelden. Zelfs als vervolgens achteraf blijkt dat een bericht nep was blijft er een residu van de onwaarheid achter bij mensen; een 'er was toch iets met...'-gevoel. Ook kan de verspreiding van nepnieuws zo snel gaan dat de effecten zich al hebben gematerialiseerd voordat het bericht kan worden ontkracht. Het tweede, en wellicht nog belangrijkere, verschil is de democratisering van de techniek. De verwachting van alle respondenten was dat de techniek binnen een jaar of twee, drie in handen zou zijn van de gewone burger en grif zou worden gebruikt. Gratis apps zijn nu al beschikbaar en die apps zouden huns inziens alleen maar beter en sneller worden. Het vervaardigen van zeer realistische deepfakes kan dan binnen een handomdraai door vrijwel iedere burger ter wereld. Daarmee zal het aantal fake-content exponentieel toenemen, zozeer zelfs dat er inderdaad een dreiging kan ontstaan dat echt niet meer van nep te onderscheiden is. De hoeveelheid nepcontent kan volgens sommigen zelfs de hoeveelheid authentieke content overstijgen. Daarmee wordt het noodzakelijk om alle content op authenticiteit te checken, terwijl dit praktisch en financieel gezien vrijwel ondoenlijk zal zijn. Dit zal onder meer een zware last leggen op de journalistiek.

Ook wordt erop gewezen dat mensen steeds meer in verwarring zullen raken. Is iets nu echt of nep? Het ene gevaar is dat nepberichten voor waar worden aangenomen. Zeker zo groot is het gevaar dat echte berichten niet meer worden geloofd. Verschillende geïnterviewden wezen erop dat deze tendens nu al, terwijl deepfaketechnologie nog in de kinderschoenen staat, aan de gang is. Ook werd erop gewezen dat een te verwachten nieuwe toepassing van deepfaketechnologie is om niet beelden te manipuleren, maar om artefacten van manipulatie achter te laten op echte beelden en geluiden, zodat detectiesoftware deze content eruit filtert en automatisch blokkeert. De verwarring over wat waar is en wat niet zal daarmee de komende jaren alleen maar toenemen.

Anderzijds wordt er ook gewezen dat de vrees voor de teloorgang van de waarheid al eeuwen bestaat en bij de introductie van iedere technologie weer opspeelt. De introductie van de drukpers stelde particulieren in staat allerhande pamfletten met opinies en halve waarheden op grote schaal te verspreiden, het internet stelde mensen in staat in bubbels te leven, de virtuele wereld zou de echte wereld op termijn vervangen. Steeds is het wel

meegevallen met die vrees. Eerder is het zo dat op de introductie van een nieuwe techniek een periode van chaos en ongeregeldheden volgt, waarna er zowel juridische, maatschappelijke als institutionele normen worden ontwikkeld om het gebruik van de technologie in goede banen te leiden.

Tot slot is van belang dat een aantal respondenten aangaf moeite te hebben met de technologie als zodanig, zelfs als die voor 'positieve' toepassing wordt gebruikt. Een politicus die een deepfake gebruikt om een minderheid in zijn land in hun taal te adresseren, terwijl hij die taal eigenlijk niet machtig is, is dat niet gewoon misleiding en kiezersbedrog? Een overleden kunstenaar die een rondleiding geeft in een museum of een historisch figuur die als deepfake geschiedenisles geeft aan scholieren, is dat niet gewoon creepy? Hoever moet de politie gaan om neppersonen in te zetten om kindermisbruik of vrouwenhandel tegen te gaan, als zij daarmee bijdraagt aan een wereld waarin nep steeds minder van echt te onderscheiden is? Het satire bedrijven door middel van deepfakes over bekende personen of naasten kan leuk zijn, maar gaat toch vrijwel altijd ten koste van een ander; zeker als die satire nauwelijks of niet meer van echt te onderscheiden is, zullen de negatieve consequenties de positieve vaak overstijgen. Anderzijds wijzen geïnterviewden erop dat deze aarzeling een conservatieve reflex is; over een aantal jaar zal dit soort toepassingen heel gewoon zijn, net zoals nu al hologrammen van overleden personen concerten geven en dergelijke concerten graag worden bezocht door jong en oud.

6.4.3 Procesrecht

Het huidige procesrecht gaat grotendeels uit van vertrouwen. De rechter gaat ervan uit dat het bewijs dat de partijen aanleveren in principe waarheidsgetrouw is, tenzij er aanleiding is om te twijfelen. Wel zijn er bij bepaalde type bewijsmiddelen algemeen bekend dat daar een onzekerheidsmarge bij geldt. Bij DNA-materiaal is dat gekwantificeerd met percentages; bij getuigen is bekend hoe onzeker het geheugen is en zal ook goed worden gekeken naar steunbewijs/feiten en omstandigheden die het verhaal van de getuige ondersteunen of ontkrachten. Maar zeker bij documenten en ander 'objectief' materiaal zal er niet worden getwijfeld, tenzij daar aanleiding toe is of een van de partijen de waarachtigheid betwist.

Als een van beide partijen een bewijsmiddel betwist dan zal in principe de partij die de waarachtigheid en juistheid van documenten in twijfel trekt aannemelijk moeten maken dat er reden is om te twijfelen. Als dat lukt zal de partij die het bewijsstuk heeft ingebracht moeten aantonen dat de waarachtigheid en juistheid wel dient te worden aangenomen door de rechter. Waar deze grenzen precies liggen en wie nu precies wat op welk punt aannemelijk moet maken valt niet eenduidig vast te stellen, zo bleek uit de interviews. Uiteindelijk benadrukken alle geïnterviewden op dit punt dat het gaat om de overtuiging van de rechter, die daarbij kijkt naar het geheel aan feiten en omstandigheden. Het gaat dus om een subjectief oordeel.

De te verwachten toename in het aantal deepfakes zal volgens de geïnterviewden leiden tot een toename in de duur van de rechtszaken en de rol van experts binnen het rechtsproces vergroten. Een partij kan vrijwel altijd stellen dat een bewijsmiddel fake is. Dit betekent dat er steeds langer dient te worden stilgestaan bij bewijsmiddelenkwesties. Bij de toename van de rol van experts, die ook al bij andere

bewijsmiddelen valt te zien, wordt gewezen op de zogenoemde deskundigenparadox, wat wil zeggen dat hoe technischer de bewijsmiddelen worden, hoe meer een rechter op het oordeel van deskundigen zal moeten stoelen, terwijl hij de kennis ontbeert om het werk van de deskundige goed te kunnen waarderen.

Ook van belang is dat er een plicht is om waarheidsgetrouwe informatie en bewijsmiddelen aan de rechter te verstrekken, maar dat het in de ervaring van de geïnterviewden maar zelden voorkomt dat zware sancties worden gekoppeld aan het aanleveren van leugenachtig bewijs. Doorgaans zal binnen het civiel recht de partij die het valselijke bewijs heeft aangedragen de zaak verliezen. Slechts in hoog uitzonderlijke gevallen wordt er bijvoorbeeld ook tuchtrechtelijk opgetreden tegen advocaten die willens en wetens vals materiaal hebben verstrekt. In interviews gehouden met internationale experts werd keer op keer gewezen op het gevaar van deepfakes in de rechtszaal. Zij zagen dit als wellicht een van de meest ingrijpende gevolgen van deepfaketechnologie. Ook opperden zij dat er wellicht met een duidelijke strafbaarstelling moet komen voor het inbrengen van deepfakes in een rechtszaak. Zij wijzen erop dat het nu al, in zaken waar mogelijk deepfakes zijn ingebracht, het voor experts vrijwel ondoenlijk is om met zekerheid vast te stellen dat materiaal is gemanipuleerd en zo ja, welk gedeelte daarvan.[400] Met de Nederlandse geïnterviewden is ook de mogelijkheid om advocaten een grotere zorgplicht te geven de revue gepasseerd, bijvoorbeeld door aan hen een zorgplicht op te leggen om ervoor zorg te dragen dat het door hen aangeleverde bewijs waarheidsgetrouw is.

Ook uit het interview met de Nederlandse Politie bleek dat het gevaar van deepfakes op bewijsmateriaal binnen de context van vervolging en handhaving zeker niet wordt onderschat. De politie investeert momenteel in systemen en procedures om het binnengekomen materiaal op waarheidsgetrouwheid te controleren. Toch is men zich daar juist bewust dat gegeven het feit dat over een aantal jaar meer dan 90% van het digitale materiaal is gemanipuleerd, het ondoenlijk zal blijken om al het materiaal op individueel niveau zorgvuldig na te lopen. Bovendien ontbreken daar de middelen voor.

Tot slot is besproken de mogelijkheid om bepaald bewijsmateriaal alleen maar in de rechtszaal toe te laten als het aan bepaalde kenmerken voldoet. De analogie werd getrokken met mails, die vrij eenvoudig te vervalsen zijn. Wellicht zal er op termijn een regel komen waaruit volgt dat mails als bewijsstuk alleen toelaatbaar zijn in de rechtbank als ze met een bepaald programma zijn verstuurd en opgeslagen, waarbij dat programma een authenticiteitsstempel afgeeft. Zo'n systeem kan mogelijk op termijn ook worden geïntroduceerd ten aanzien van video en ander materiaal.

6.4.4 Regulering

Verschillende reguleringsopties zijn in de interviews de revue gepasseerd.

Allereerst wordt er gedacht aan bewustwording. Een groot probleem dat algemeen werd onderschreven betreft de onbekendheid met deepfakes onder het grootste deel van de bevolking. Burgers zijn zich er niet van bewust hoe eenvoudig zij kunnen worden misleid; misschien nog kwalijker, personen die in hun professionele hoedanigheid te maken krijgen met deepfakes, zoals journalisten en rechters, denken vaak dat zij wel in staat zullen zijn om echt van nep te onderscheiden. Ten onrechte. Ook is het van

belang om een publiekscampagne te starten waarbij sociale normen worden neergelegd. Dat die wel degelijk succesvol kunnen zijn blijkt onder meer uit de anti-rookcampagne die in diverse landen significante effecten heeft gehad. Hierbij valt met name te denken aan deepfakes voor non-consensual porn, zeker, maar niet uitsluitend, als die is gericht op tieners. Jongeren maken dergelijke filmpjes nu wellicht voor de grap of om stoer te doen, zonder dat zij zich echt bewust zijn van de gevolgen daarvan voor tienermeisjes. Het moet ondubbelzinnig duidelijk worden dat dergelijke fake-filmpjes niet kunnen. Sancties kunnen hierbij een afschrikwerkende of signaalfunctie hebben.

Vervolgens benadrukte een aantal geïnterviewden dat het recht nu met name bescherming biedt aan welgestelden. Zij hebben het geld en de middelen om lange en complexe rechtszaken te voeren tegen platformen en gebruikers daarvan die over hen deepfakes verspreiden, terwijl gewone burgers vaak de technische expertise, de financiële middelen en de kennis omtrent het rechtssysteem ontberen om dergelijke, soms jarenlang durende procedures, aan te gaan. De overheid zou wat dat betreft dus meer moeten inzetten op de bescherming van burgers, bijvoorbeeld vrouwen die slachtoffer worden van deepfake porno. Dat kan op een aantal manieren. Een daarvan is het aanstellen van een speciale Ombudsman, of het vergroten van de bevoegdheden en de middelen van de bestaande Ombudsman, waar burgers hun zaken kunnen aandragen en die namens hen gesprekken en waar nodig procedures aanspant tegen internetplatforms.

De meeste respondenten wezen op de problemen van handhaving van regulering in deze context om een aantal redenen. Ten eerste is niet van tevoren te voorspellen hoe deepfaketechnologie zal worden ingezet en of die inzet schade zal veroorzaken. Ten tweede is er in de digitale omgeving een complex web aan partijen bij het genereren en verspreiden van een deepfake. De ontwikkelaar van de technologie, de App-producent, de App-store, de burger die een filmpje maakt, de site waarop het fake bericht wordt verspreid, de andere gebruikers die het bericht downloaden of verder verspreiden, zoekmachines en andere indexen die verwijzen naar de content, etc. Op welke partij precies welke verplichting dient te worden gelegd is niet op voorhand duidelijk. Ten derde spelen er in verschillende jurisdicties verschillende regels, zodat de maker of verspreider van een deepfake in zijn eigen land niet onrechtmatig handelt, maar wel in het land waar het slachtoffer woonachtig is. Ten vierde is het lastig om regels af te dwingen over landsgrenzen heen.

Toch werd unisono benadrukt dat regulering essentieel is. Een geïnterviewde vergeleek het met spam. Als spam niet zou worden tegengegaan op diverse manieren dan zou e-mailen als zodanig ondoenlijk worden, het hele systeem zou onbruikbaar worden. Zo geldt het ook voor deepfakes. Het hele internet zou onbruikbaar dreigen te worden als daar inderdaad nep niet meer van echt te onderscheiden is. Welke vorm van regulering dient te worden aangenomen, daarover werd verschillend gedacht. Alhoewel iedereen het erover eens was dat consumententoepassing en -gebruik van deze technologie de minste potentiële waarde vertegenwoordigde, meende een flink aantal geïnterviewden, met name uit de Verenigde Staten, dat het verbieden van deze toepassing of apps bestemd voor de consumentenmarkt te ver

zou gaan. Anderen vonden dat hier in ieder geval over moest worden nagedacht en wezen daarbij op bepaalde apps, zoals de Deepnude app, die er slechts en uitsluitend op lijkt te zijn gericht om kwaadaardige en schadelijke content te produceren. Zij wezen erop dat non-consensual porn nu al in alle jurisdicties ter wereld strafbaar is; het gaat niet om de juridische inkadering van deze toepassing, maar om de handhaving daarvan. Tot slot wezen zij erop dat het gevaar dat de waarheid steeds moeilijker vast te stellen is met name is gekoppeld aan de hoeveelheid deepfakes.

Overeenstemming was er ten aanzien van het feit dat er flink geïnvesteerd moest worden in deepfake-detectietechnologie, ook al wezen de geïnterviewden met een technische achtergrond erop dat dit alleen een eerste bouwsteen kon zijn. Daarnaast werd het watermerken van fake-content wenselijk bevonden, maar was hierbij ook duidelijk de vraag of kwaadwillenden zich hieraan zouden houden en hoe dit moest worden gecontroleerd en overtredingen gesanctioneerd. Breed gedragen was ook het idee dat met name sociale media en internetplatforms een cruciale rol spelen, omdat zij essentieel zijn bij de brede verspreiding van deepfakes. Welke vorm dergelijke regulering diende te krijgen, daarover verschilden de meningen. Sommigen wilden primair op zelfregulering inzetten, anderen op wetgeving.

Tot slot werd geopperd om meer in te zetten op publiek gefinancierde media, die een centralere rol kunnen spelen bij het vergaren, analyseren en op waarheid checken van informatie. Alhoewel de gedachte vaak is dat door de opkomst van het internet de publieke omroepen minder van belang zijn geworden is dat niet of slechts zeer gedeeltelijk waar. Traditionele publieke radio en televisie vormen nog steeds een van de belangrijkste informatiebronnen voor de gemiddelde burger en ook hun internetpagina's worden, juist in de Corona-crisis, vaak geraadpleegd en ook zijn in de meeste landen juist de publieke omroepen zeer actief met open access online content en zijn de podcast platforms van publieke omroepen en traditionele media leidend. Daarom zou de overheid in het tijdperk van desinformatie meer kunnen investeren in publieke omroepen of kwaliteitsmedia kunnen ondersteunen, om zo in ieder geval eilanden van authenticiteit en betrouwbaarheid op het internet te genereren.

6.5 Conclusie

In dit hoofdstuk is een algemene landeninventarisatie gegeven en een samenvatting geboden van twee landenstudies die voor dit onderzoek zijn uitgevoerd en in het geheel in de bijlagen te vinden zijn (paragraaf 10.1). Ook is de neerslag gegeven van de interviews die voor deze studie zijn gedaan (volledige interviewverslagen zijn te vinden in de bijlagen bij dit rapport, paragraaf 10.2). Hieruit volgt een diversiteit aan reguleringsopties, benaderingen en overwegingen, die als basis zullen dienen voor de reflecties (hoofdstuk 7) en de reguleringsopties voor de Nederlandse wet- en regelgever (hoofdstuk 8).

Wat betreft de landen valt een varieteit te zien aan thema's die worden geadresseerd, zoals misinformatie, deepfake porno en mogelijke beïnvloeding van het democratisch proces. Ook wat betreft de aanpak is een breed scala aan mogelijkheden te zien; zo wordt er geopteerd voor de strafrechtelijke route, staat in andere jurisdicties de civielrechtelijke route centraal

en wordt in sommige landen gekozen voor technologisch georiënteerde maatregelen. Het primaire zwaartepunt ten aanzien van het toezien op de naleving van de maatregelen wordt belegd bij diverse partijen, zoals het Openbaar Ministerie, internet intermediairs en de burger.

In de interviews wordt onder meer gewezen op de moeilijkheid om deepfakes al dan niet via technische weg te onderscheiden van niet gemanipuleerde content en wordt de verwachting uitgesproken dat meer dan 90% van de online content op termijn gemanipuleerd zal zijn. Ook wordt gewezen op de grote maatschappelijke gevolgen van deepfakes en op de diverse obstakels voor een effectieve regulering van deepfakes.

Voetnoten hoofstuk 6

- 383 <https://www.theguardian.com/australia-news/2021/feb/22/australian-government-ready-to-pursue-facebook-and-twitter-if-misinformation-code-doesnt-work>.
- 384 <https://digi.org.au/disinformation-code/>.
- 385 <https://www.esafety.gov.au>.
- 386 <https://cyber.gc.ca/sites/default/files/publications/tdp-2019-report_e.pdf>.
- 387 <https://www.eyerys.com/articles/timeline/german-law-allows-authorities-use-deepfakes-children-catch-online-predators?page=12#event-a-href-articles-timeline-facecom-facebookfacecom-for-facebook-a>.
- 388 <https://thenextweb.com/news/german-investigators-to-use-deepfake-images-of-child-sexual-abuse-to-bust-online-predators>.
- 389 <https://news.microsoft.com/de-de/artificial-intelligence-to-make-great-strides-in-the-fight-against-child-pornography/>.
- 390 <https://dserver.bundestag.de/btd/19/156/1915657.pdf>.
- 391 <https://www.researchgate.net/publication/338740911_Desinformation_aufdecken_und_bekampfen_Interdisziplinare_Ansatze_gegen_Desinformationskampagnen_und_fur_Meinungspluralitat>.
- 392 Senate P.S. Res. No. 188, 4 Nov. 2019.
- 393 <http://legacy.senate.gov.ph/lisdata/3176028609!.pdf>.
- 394 <https://www.legifrance.gouv.fr/loda/id/JORFTEXT000037847559/>.
- 395 <https://www.researchgate.net/publication/345809619_deepfakes_and_the
- 396 <http://164.100.24.220/loksabhaquestions/annex/172/AU2613.pdf>. Zie verder: Sections 67, 67A, 67B, 67C and 67D of the Information Technology Act, 2000. < https://www.deccanherald.com/national/north-and-central/deepfake-videos-were-used-for-the-first-time-in-india-by-bjp-report-806669.html>. <https://www.hindustantimes.com/mumbai-news/bombay-hc-seeks-information-from-i-b-ministry-on-ai-bot-that-turns-photos-into-nudes/story-U8B1FkeYCFw3WRjYl41P3L.html>. <https://www.ndtv.com/india-news/bombay-high-court-seeks-info-from-centre-on-ai-bot-that-turns-photos-into-nudes-2313802>. <https://blogs.lse.ac.uk/southasia/2020/05/21/deepfakes-in-india-regulation-and-privacy/>.
- 397 <https://www.wilsoncenter.org/blog-post/ukraines-new-media-laws-fighting-disinformation-or-targeting-freedom-speech>.
- 398 <https://sso.agc.gov.sg/Acts-Supp/18-2019>.
- 399 <https://commonslibrary.parliament.uk/research-briefings/cbp-8743/>.
- 400 Zie ook: <https://www.dailymail.co.uk/news/article-9592117/Cops-accused-woman-creating-deepfake-images-never-evidence.html>

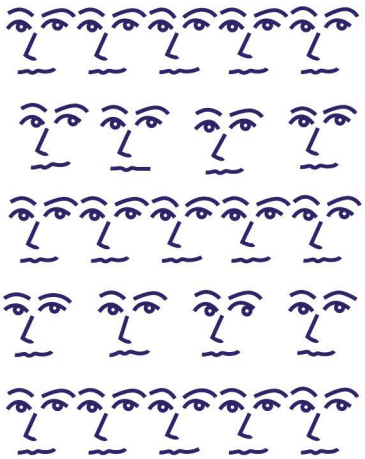
Bron

Sociale media

Nieuws

Website politieke partij

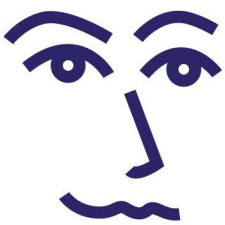
Verzamelen/verwerken



Toepassing

Satire

Ontwrichten



Procederen

Een andere context betekent dat andere voorwaarden van kracht zijn (AVG). Dat iets op het internet staat betekent niet dat je het zomaar kunt gebruiken in een ander context. Dit geldt zowel voor brondata, verwerking van de data en de toepassing in bijvoorbeeld een deepfake of andere content bewerking.

7. Reflecties

In dit hoofdstuk zullen de belangrijkste inzichten worden besproken, almede het algemene kader waarbinnen de reguleringsopties (hoofdstuk 8) moeten worden bezien. Hierbij zal gebruik gemaakt worden van en dus overlap bestaan met de vorige hoofdstukken uit dit rapport. Paragraaf 7.1 geeft de reflecties op de technische mogelijkheden en de beperkingen van de deepfaketechnologie. Paragraaf 7.2 staat stil bij de gevaren en risico's van deepfakes, terwijl paragraaf 7.3 stilstaat bij de belangen en de kansen van de inzet van deepfakes in horizontale verhoudingen. Paragraaf 7.4 bediscussieert het punt dat techniek, zoals deepfaketechnologie, neutraal zou zijn. Paragraaf 7.5 bespreekt het verschil en de overlap tussen de vragen die door deepfaketechnologie worden opgeroepen en reeds bestaande fenomenen en ontwikkelingen en vragen die daarmee samenhangen. Tot slot zoomt paragraaf 7.6 in op de handhaafbaarheid van de bestaande regels ten aanzien van deepfakes in het moderne medialandschap.

7.1 Technische mogelijkheden en beperkingen

Deepfaketechnologie stelt een gebruiker in staat om bestaand materiaal samen te voegen en te manipuleren of nieuw materiaal te genereren. Daarbij kan het gaan om video, audio of tekst, maar ook breder, om elk ander type signaal of informatie. Er zijn verschillende toepassingsmogelijkheden. De meest simpele versie is om een bestaande video van een persoon te nemen en daar het gezicht van een ander op te monteren. In meer geavanceerde toepassingen kunnen ook bijvoorbeeld gezichts- of lichaamskenmerken van twee of meer personen worden samengevoegd. Daarnaast kan een bestaand beeld, video- of audiofragment zo worden gemanipuleerd dat de geportretteerde er anders uit komt te zien, andere handelingen lijkt te verrichten en/of andere woorden lijkt te spreken. Tot slot is het mogelijk om nieuwe personen te genereren.

Alhoewel de deepfaketechnologie pas een klein aantal jaar oud is, zijn de technische mogelijkheden in rap tempo vooruitgegaan. De verwachting is dat deze ontwikkelingen zich in de toekomst verder zullen doorzetten. Nu al is een deepfake die door een professioneel team is gemaakt met het blote oog niet van echt te onderscheiden. De verwachting is dat deze geavanceerdere technieken binnenkort ook op de consumentenmarkt zullen verschijnen. Hierdoor is het voor burgers mogelijk om middels een gratis app fake video's of geluidsfragmenten van zichzelf of anderen te genereren, als zij toegang hebben tot beeld- of geluidsmateriaal. Dit is een kwestie van seconden of minuten. Daarbij hoeft de burger in kwestie niet per definitie toegang te hebben tot materiaal dat lijkt op het beoogde eindproduct; zo bestaan er nu al apps die burgers in staat stellen om een beeld van een ander waarop die gekleed is te gebruiken om een fake-naaktbeeld van diegene te genereren.

De verwachting is dat het gebruik van deepfaketechnologie de komende jaren een grote vlucht zal nemen. Daarbij speelt dat steeds meer technieken, of misschien beter gezegd, informatietechnologie per definitie, de waarheid manipuleert. Dat gaat vaak om relatief kleine manipulaties: fotocamera's die rode ogen genereren, video-beldiensten die de huid van een persoon egaler laten lijken, audio die middels compressie een deel van zijn kwaliteit

en een aantal van de hogere geluidsregisters verliest. Toch kunnen deze kleinere manipulaties van groot belang zijn, bijvoorbeeld bij de identificatie van een verdachte of bij een medisch online-consult. De verwachting is dat de synthetische werkelijkheid, dat wil zeggen materiaal dat gedeeltelijk is gemanipuleerd door digitale technieken of daardoor is gegenereerd, een grote vlucht zal nemen. Experts 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 bijna onmogelijk om met het blote oog vast te stellen of een video of ander materiaal een deepfake is of niet – hierbij maakt het niet uit of de persoon alvorens het zien van de deepfake bekend is met het fenomeen of niet[401] – 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. Mede doordat detectietechnologie gebaseerd is op de deepfaketechnologie, is de laatste technologie altijd een stap voor. Bovendien wijzen zij erop dat ook het omgekeerde probleem zal ontstaan: het is vrij eenvoudig om met deepfaketechnologie op een bestaand en niet gemanipuleerd materiaal sporen (artefacten) van manipulaties achter te laten. De detectie-technologie zal het materiaal dan aanmerken als fake en 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%. Dit roept nieuwe vragen op, bijvoorbeeld voor de media en binnen het rechtsproces, zoals: bij welk betrouwbaarheidspercentage mag materiaal als bewijsmiddel dienen in de rechtszaal en ligt dat percentage hoger binnen het strafrecht dan binnen het burgerlijk recht? Dient de overheid een hoger waarheidspercentage als standaard te hanteren dan het bedrijfsleven en moet de publieke omroep een hoger waarheidspercentage hanteren dan commerciële nieuwsorganisaties en zo ja, welk percentage dient dan te worden gehanteerd?

De beste strategie om deepfakes te ontdekken is volgens experts niet door middel van tegen-technologie, maar door middel van een menselijke inschatting naar de waarheid op basis van contextuele informatie: is dit iets wat deze persoon normaal gesproken zou zeggen, sluit dit aan bij andere informatie omtrent dit onderwerp of deze persoon, zijn er andere bronnen die hetzelfde berichten, etc. Toch is de nieuwste ontwikkeling waarop zij wijzen precies hierop gericht. AI wordt ingezet om hele fake omgevingen in te richten: niet alleen een deepfake video, maar ook fake nieuwswebsites die daar een reportage over maken, fake Twitter-accounts die met elkaar in discussie gaan over het filmpje, fake Instagram-accounts waar memes van het filmpje worden gegeneerd, Wikipagina's die automatisch worden aangepast of aangemaakt over dit onderwerp, fake-journaalitems die over de ophef en discussie op Twitter berichten, etc. Zo ontstaat een hele omgeving van nepinformatie, waardoor het zowel voor mens als algoritme steeds lastiger wordt om echt van nep te onderscheiden.

7.2 Grote gevaren en maatschappelijke vragen

Dit rapport heeft in hoofdstuk 2 de mogelijke positieve en negatieve kanten van deepfaketechnologie besproken, de kansen en risico's, zoals dat dan in het beleidsjargon heet. Het is meer in het algemeen de vraag of dit altijd het beste kader is om beleidsvraagstukken in te bespreken. Moord of verkrachting, om maar een extreem voorbeeld te geven, bespreek je niet in termen van het nadeel van het slachtoffer versus de voordelen voor de dader – dat is gewoon fout. Een utilitaire belangenafweging geeft de sug-gestie dat niets principieel goed of fout is, maar dat het gaat om een inschatting over de te verwachten voor-en nadelen. Dat kan op zijn plaats zijn bij economische vraagstukken, bij juridische en morele vaak niet. Bovendien is de suggestie die bijna automatisch van zo'n kader uitgaat: de wetgever moet zorgen dat de kansen optimaal worden benut en de risico's worden beperkt of geminimaliseerd. Ook dit is niet altijd de beste conclusie. Tot slot kan zo'n kader geen antwoord bieden op de grotere, onderliggende vragen. Sommige experts voorspellen bijvoorbeeld een wereld van Singularity, waarin mens en machine een worden, en anderen een wereld waarin de mens in zijn geheel wordt vervangen door de machine. Wat is dan precies de kans en wat is het risico bij zo'n ontwikkeling? En preciezer, de vraag is niet naar concrete kansen en risico's van een dergelijke ontwikkeling, de vraag is in wat voor een wereld we willen leven.

Deze meer algemene aarzelingen gelden zeker ook in het kader van deepfakes. Het is duidelijk dat de mogelijke gevaren niet alleen de mogelijke voordelen in aard en ernst overstijgen, ook is een aantal van de maatschappelijke gevolgen of risico's simpelweg van een andere orde.

Dit rapport ziet primair op de inzet van deepfakes in horizontale verhoudingen, dat wil zeggen in burger-burger relaties, waarbij de ene burger over de andere een deepfake maakt. De democratisering van de deepfake-techniek heeft mogelijk tot gevolg dat er zoveel nepcontent op het internet verschijnt dat waarheid en fictie steeds meer door elkaar gaan lopen. Dit gecombineerd met het feit dat op termijn meer dan 90% van het digitale materiaal gemanipuleerd zou kunnen zijn, kan een zeer ontwrichtend effect hebben op de samenleving. Daarbij valt te denken aan een toenemende polarisatie tussen groepen die steeds meer in hun eigen werkelijkheid gaan geloven. Ook is het voor media vrijwel ondoenlijk om alle content stelselmatig op betrouwbaarheid te controleren, zowel in termen van tijd als middelen. Bovendien is het onvermijdelijk dat steeds meer materiaal door de mazen van het net heen glipt, zodat officiële kanalen inderdaad fake-nieuws zullen brengen en kan een mogelijk nadeel zijn dat als betrouwbare media strikte regels en procedures hanteren dat zij altijd twee stappen achter media aanlopen die wel de sensationele (eventueel fake) nieuwberichten direct plaatsten.

Ook kan door de opkomst van deepfakes de rechtstaat onder druk komen te staan. Ten eerste kunnen juridische processen langer duren, omdat partijen altijd kunnen beweren dat tegen hen geleverd bewijs nep en gefabriceerd is. Dit kan dan nader onderzoek vergen en een discussie in de rechtszaal opleveren tussen diverse experts die elk zo hun eigen visie hebben om de eventuele authenticiteit van het materiaal. Ten tweede is het gevaar dat de rechter content onterecht voor waar zal aannemen en dit tot een onterechte beslissing leidt. Ook het omgekeerde, namelijk dat de rechter meent dat bepaald

materiaal (mogelijk) fake is, terwijl dat niet het geval is, kan onwenselijke gevolgen hebben. Ten derde kan een veroordeelde, na een rechtelijke uitspraak, altijd publiekelijk zijn onschuld volhouden door te beweren dat de rechter in een fake-bericht is getrapt. Ten vierde kan bij bepaalde delicten de suggestie die uitgaat van een deepfake al genoeg zijn voor publieke verontwaardiging en een publieke veroordeling, ook al komt het niet tot vervolging of wordt later bekend dat content nep was.

Ook de democratie als zodanig kan onder druk komen te staan. Voor dit gevaar is momenteel reeds de meeste aandacht. Hierbij wordt met name gewezen op incidenten waarbij buitenlandse mogendheden, en dan met name Rusland, fakeberichten en trollen lijkt in te zetten om democratische processen te beïnvloeden. Een aantal landen en meerdere staten van de Verenigde Staten, hebben reeds wetgeving aangenomen om zich hiertegen te wapenen. Ook is duidelijk dat belangengroepering binnen de landsgrenzen deepfakes inzetten om hun politieke wensen over het voetlicht te brengen of door hun gesteunde kandidaten in een goed daglicht te stellen. Daarnaast wijzen experts, onder meer in de voor deze studies gehouden interviews, erop dat staten ook concrete, voor hen relevante, besluitvorming in andere landen trachten te beïnvloeden door de verspreiding van nepnieuws, bijvoorbeeld in de Global South. Door deze toepassingen kunnen deepfakes een potentieel zeer ontwrichtende werking hebben op het democratisch proces en de internationale rechtsorde.

Tot slot moet er ook rekening worden gehouden met de gevolgen van deepfakes voor de sociale veiligheid en maatschappelijke positie van vrouwen en jonge meisjes. Veel van de deepfakes die momenteel worden gemaakt betreffen seksuele afbeeldingen of video's van vrouwen. Dit heeft ten gevolg dat het vrouwenlichaam nog meer dan nu wordt geseksualiseerd, dat onrealistische schoonheidsidealen kunnen worden bevestigd en dat vrouwen worden gestigmatiseerd. Slutshaming en misogyne opmerkingen zijn nu al aan de orde van de dag, zowel offline en zeker ook online, iets wat deepfaketechnologie alleen maar zal uitvergroten. Nu nog worden veel privéopnames van seksuele handelingen openbaar gemaakt door een ex-partner – de zogenoemde wraakporno. Met deepfaketechnologie in de hand is een dergelijke reële opname niet meer nodig; iedere puberende jongen kan een fake-pornofilmpje van het mooiste meisje uit de klas genereren en verspreiden op sociale netwerken, al dan niet in besloten kring. Niet alleen jonge meisjes kunnen slachtoffer worden. Ook vrouwelijke beroemdheden en politici lopen gevaar en kunnen het slachtoffer worden van een pornografische deepfake, waardoor de geloofwaardigheid van deze vrouwen wordt ondermijnd. Deze vrouwen kunnen zodoende reputatieschade oplopen, wat bijvoorbeeld voor politici als gevolg kan hebben dat zij zullen (moeten) aftreden.

Experts wijzen erop dat de wetenschap dat een filmpje deepfake is maar van beperkt belang is. De sociale consequenties van een porno-filmpje voor een opgroeiend meisje kunnen aanzienlijk zijn, ook al weet de groep dat het een deepfake betreft. Ook kan het zien van een dergelijk filmpje het zelfbeeld van de vrouw in kwestie aantasten. Ook al weet ze dat het materiaal nep is, toch kan het bekijken van jezelf terwijl je allerhande expliciete handelingen

verricht een negatieve impact hebben op je zelfvertrouwen en zelfwaarde. Dit punt – dat ook al is het bekend dat bepaalde content fake is, de gevolgen er niet minder om zijn - geldt bijvoorbeeld ook bij fakenieuws. Berichten die op het eerste gezicht zeer dubieus lijken worden maar al te graag omarmd en gedeeld binnen groepen waarbinnen het bericht aansluit op hun wereldbeeld en/of politieke overtuiging (dat dit zelfs zonder deepfakes al gebeurt blijkt uit het grote percentage van Amerikaanse kiezers die in één of meerdere fabels van Trump omtrent zijn verkiezingsnederlaag gelooft). Ook als later het bericht wordt ontkracht, dan nog houdt zo'n groep vaak vol dat het specifieke bericht dan wel nep was, maar dat de onderliggende waarheid die daarvan uitging wel degelijk klopt. Bij nepnieuws dat onder het grotere publiek wordt verspreid, maar later wordt ontkracht, is de vrees dat het aanvankelijke (vaak sensationele) fake-bericht significant meer aandacht zal genereren dan de latere nuance of ontkrachting van dat bericht. Bovendien blijft er dan vaak alsnog een 'er was toch iets met...' gevoel achter.

Tot slot stelt deepfaketechnologie ook de vraag: willen we dit eigenlijk wel? Deze vraag kan niet wetenschappelijk beantwoord worden, maar is wel van belang als het parlement en de regering hierover debatteren en eventuele reguleringsopties bespreken. Een voorbeeld is de inzet van deepfaketechnologie voor het weer tot leven wekken van dode personen. Een overleden kunstenaar kan bijvoorbeeld een rondleiding geven in een museum; Napoleon kan op middelbare scholen geschiedenisles geven; familieleden kunnen zien hoe hun betovergrootmoeder er waarschijnlijk uit had gezien en geklonken; een overleden persoon kan op zijn eigen begrafenis spreken; een overleden zanger kan concerten geven (Elvis is back); en partners kunnen middels een deepfake van hun overleden echtgenoot in contact blijven en doen alsof die er nog is. Dit wordt vaak als positieve toepassingen beschreven in de literatuur. Maar sommige experts vragen zich af of dat wel zo is. Draagt het college laten geven door historische figuren aan schoolkinderen bijvoorbeeld niet bij aan een gewenning aan de post-truth world; leidt het blijven converseren met een overleden partner niet juist tot psychische problemen, omdat de overgebleven partner nooit aan echte rouwverwerking toekomt; wilde betovergrootmoeder eigenlijk wel tot leven worden gewerkt? Vergelijkbare vragen spelen ook bij andere deepfake-toepassingen, zoals bij het inzetten van volledig fictieve personages door de politie bij de opsporing van kinderpornonetwerken, vrouwenhandelaren en georganiseerde misdaad. Hoe ver kan en mag de politie daarin gaan? En een politicus die een minderheid in zijn land in hun eigen taal toespreekt, terwijl hij die taal niet daadwerkelijk machtig is; is dat wenselijk uit het oogpunt van inclusie of een vorm van kiezersbedrog?

7.3 Beperkte belangen

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.

(Uit de meeting van de Tweede Kamer met een nep-vertegenwoordiger van Navalny en het incident waarbij welgestelde Fransen aan een nep-minister grote sommen geld overmaakten voor een geheim defensieproject blijkt overigens dat deepfaketechnologie daar niet eens een vereiste voor is. Wel is duidelijk dat deepfakes dit fenomeen alleen maar zullen vergroten.) 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.

Positieve toepassingen zijn er ook. Daarbij wordt onder meer gewezen de eerdergenoemde mogelijkheden voor de politie om fakes in te zetten bij de infiltratie van criminele netwerken, om getuigen te anonimiseren en om bijvoorbeeld door fake-kinderporno pederasters op te pakken. Fake-kinderporno kan ook worden ingezet voor de behandeling van pedofielen.[402] Daarnaast zijn er medische toepassingen, bijvoorbeeld voor mensen die een problematisch zelfbeeld hebben, en kunnen er medische visualisaties van het menselijk lichaam of een onderdeel daarvan worden gemaakt. Een plaats delict kan middels een deepfake zeer realistisch worden nagebootst, een realistische deepfake avatar kan figureren in een game, een keukengigant kan een deepfake impressie van het huis met het nieuwe kookeiland geven, een deepfake van een acteur kan gevaarlijke stunts uitvoeren en vrouwen die in de seksindustrie werken kunnen een deepfake inzetten om bepaalde activiteiten te verrichten. Deepfakes van dode personen, zoals hierboven aangestipt, kunnen worden ingezet voor onder meer rondleidingen in musea, het geven van concerten, het voordragen van een geschiedenisles of het figureren in een film.

Ook kunnen deepfakes worden ingezet voor het vervormen van alle bestaande films, zodat niet alleen de audio wordt vertaald naar het Engels, maar ook de videoafbeeldingen van de lippen van de acteurs zo worden vervormd dat het lijkt alsof zij inderdaad Engels spreken. Deze toepassing kan ook worden ingezet door politici die minderheden willen toespreken, BN´ers die oproepen tot een goed doel en in zakelijke gesprekken, bijvoorbeeld tussen Nederlandse en Chinese werknemers, waarbij wat zij in hun eigen taal zeggen real time wordt vertaald, maar wederom hun lipbewegingen daarop worden aangepast. Tot slot kan deepfaketechnologie ook worden ingezet door online winkels, bijvoorbeeld door een deepfake van een potentiële klant kleiding, brillen of andere aankopen op het lichaam te laten dragen, zodat de klant zich daar een goed oordeel van kan vormen.

Bij deze brede verzameling van mogelijke positieve use cases van deepfaketechnologie valt een ding op: het zijn veelal toepassingen binnen professionele relaties, zoals tussen klant en bedrijf, 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 een veelvoorkomende toepassing van deepfaketechnologie in burger-burgerrelaties geïdentificeerd en dat is de inzet voor satire. Deze satire kan zich op allerlei mensen met allerlei maatschappelijke posities richten. Een deepfake van een jarige partner waarin zij naast haar favoriete idool op het podium staat; de onaardige buurvrouw die in de buurtapp middels een fake-video wordt getoond met een hoge puntmuts en vliegend op een bezemsteel; een deepfake van een acteur, bijvoorbeeld van Nicolas Cage,

waardoor het lijkt alsof hij in werkelijk alle films heeft gespeeld; een deepfake van de Koning waarin hij een neptoespraak houdt; een deepfake NOS-journaal met allerhande frivole inhoud; politici die woorden in de mond worden gelegd; etc. Daarnaast zullen andere toepassingen zich ontwikkelen in de privésfeer, zoals, maar niet beperkt tot, het spreken via Zoom met verre familieleden waarbij automatisch wordt vertaald en gelipsynchroniseerd; een videoboodschap van oma die haar stem kwijt is maar via deepfake toch kan spreken; enz.

7.4 Techniek is niet neutraal

Een debat dat ook bij deepfaketechnologie speelt is in hoeverre techniek neutraal is. Opmerkelijk is hoe vaak in de interviews die zijn gehouden voor deze studie, maar ook in publieke discussiebijeenkomsten over deepfaketechnologie, naar voren wordt gebracht dat technologie zelf neutraal is. Daaraan wordt dan gekoppeld de gedachte dat een technologie op zich niet goed of fout is, maar dat het gebruik daarvan goede of slechte gevolgen en toepassingen kent. Ergo: deepfaketechnologie dient niet als zodanig aan banden te worden gelegd, de specifieke kwaadaardige toepassingen dienen te worden geadresseerd. Het is een visie op techniek die nog het meest in de Verenigde Staten is terug te horen en zijn meest bekende slogan kent in NRA's credo: guns don't kill people, people kill people. Het is van belang om kort een aantal nuances te plaatsen, zonder het hele techniekdebat hier over te doen.

Techniek is niet neutraal, het is ontwikkeld; het is ontwikkeld door de mens en met een bepaald doel. Uit dat doel volgt ook de meest waarschijnlijke gebruikstoepassing. Er zijn voorbeelden bekend van technieken die zijn ontwikkeld voor het ene doel, maar vervolgens uitermate geschikt blijken voor een ander doel, maar dat is de uitzondering. Een keukenmesje is bijvoorbeeld zo ontwikkeld dat die geschikt is voor het snijden van groenten en fruit. Andere objecten zijn niet of nauwelijks geschikt voor het snijden van groenten en fruit. Een vuurwapen is ontwikkeld voor het verwonden of doden van een levend wezen. Uiteraard kan ook een keukenmesje worden gebruikt voor het doden en verwonden van een levend wezen, maar makkelijk is dat niet. Een vuurwapen kan ook gebruikt worden voor andere doeleinden dan voor het doden of verwonden van een levend wezen, maar die alternatieve gebruikstoepassingen zijn beperkt.

Uit het ontwerp van een techniek volgt derhalve een of meerdere gebruikstoepassingen. Dat valt ook te zien aan het percentage van de gebruikstoepassingen. Meer dan 99% van de gevallen waarin een keukenmesje wordt ingezet betreft het het snijden van groenten en fruit, veruit de meeste gevallen dat een vuurwapen wordt gebruikt betreft het het doden of verwonden van een levend wezen, het dreigen daarmee of het oefenen daarvoor. Niet alleen is het zo dat technieken kunnen worden gebruikt om bestaande doelen te verwezenlijken, ze maken ook nieuwe doelen en toepassingsmogelijkheden mogelijk.[403] Het op afstand doden van duizenden mensen in een klap was voor de introductie van het kernwapen ondenkbaar; het live communiceren met personen aan de andere kant van de wereld was voor het internet ondenkbaar; etc. Iedere techniek verandert zo niet alleen de mogelijkheden, maar ook de wereld als zodanig waarin de techniek een rol speelt. De wereld vormt zich als het ware naar de techniek, de

techniek vormt na verloop van tijd een natuurlijk onderdeel van de samenleving.

Natuurlijk is het nog steeds zo dat voor iedere techniek, ook de potentieel meest gevaarlijke, er ook positieve toepassingen te bedenken zijn. Zo zou het hebben van een atoomwapen een afschrikwekkende werking kunnen hebben. Vuurwapens kunnen worden gebruikt als museumstuk en kunnen worden gebruikt ter zelfverdediging. Toch is dat niet de meest voor de hand liggende en meest veelvoorkomende gebruikstoepassing, nog los van de onbedoelde toepassingen – zo zijn er in de Verenigde Staten veel doden te betreuren omdat een vuurwapen per ongeluk is afgegaan. Daarom is er in principe een verbod op het ontwikkelen van kernwapens en is het bezit van vuurwapens in Europa aan specifieke beperkingen gebonden. In principe mogen daar alleen gezagsdragers met bijbehorende training mee rondlopen, zoals politieagenten en militairen, en voor burgers zijn er in specifieke gevallen uitzonderingen, zoals voor schietclubs en momenten van geautoriseerde jacht.

Dit debat is relevant in het kader van deepfakes 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 dienst toestemming. Daarbij moet worden opgeteld het gebruik van deepfaketechnologie voor fraude, misleiding en het verspreiden van schadelijk nepnieuws. De inzet van deepfaketechnologie, de naam zegt het al, wakkert bovendien per definitie de vervaging tussen feit en fictie aan. Zelfs al worden deepfakeberichten later gecorrigeerd of wordt er zelfs direct al bij een filmpje vermeld dat het een deepfake betreft, dan nog zal dit de verwarring over wat echt is en wat nep vergroten.

Voor deze studie geïnterviewde experts geven bijvoorbeeld aan dat de verwarring over wat echt is en wat nep nu al zichtbaar is, terwijl de techniek nog maar in de kinderschoenen staat. Niet alleen wordt nepberichtgeving voor waar gehouden, met alle gevolgen van dien, ook twee andere scenario's zijn daarbij van belang. Allereerst het scenario, zoals eerder geschetst, waarin niet met zekerheid is vast te stellen of content fake is of niet en zo ja, welke gedeelte dan. Een van de geïnterviewden gaf bijvoorbeeld aan dat in een rechtszaak, waarin eerst onomstotelijk vast leek te staan dat een vrouw concurrenten van haar dochter als cheerleader schade had berokkend door fake-video's over hen te verspreiden waarin zij allerhande moreel laakbaar gedrag leken te vertonen, gevraagde getuige-deskundigen niet met zekerheid vast konden stellen of de video's inderdaad waren gemanipuleerd en zo ja, welke delen daarvan. Ten tweede dat mensen gaan twijfelen aan authentieke berichtgeving. Een andere geïnterviewde gaf het voorbeeld van iemand in haar omgeving die eenmaal in een deepfake was getrapt en nadien vrijwel alle berichten met niet direct voor de hand liggende inhoud wantrouwde.

Deepfaketechnologie draagt dus per definitie bij aan een vervaging van fictie en werkelijkheid; veruit de meeste toepassingen zijn kwaadaardig en ongewenst; en ook zijn er specifieke apps, zoals Deepnude, die lijken te zijn ontwikkeld voor kwalijke toepassingen. Dit alles betekent niet per definitie dat deepfaketechnologie slecht is en alleen maar voor slechte toepassingen kan en zal worden gebruikt. Wel dient in het debat omtrent de regulering van de technologie de

vraag te worden behandeld hoe ervoor kan worden gezorgd dat de deepfaketechnologie niet langer in veruit de meeste gevallen voor kwaadaardige toepassingen wordt ingezet, maar voor wenselijke toepassingen.

7.5 Oud en nieuw

Tegelijkertijd is van belang om te benadrukken dat deepfakes tot nu toe worden ingezet op een wijze die aansluit bij maatschappelijke tendensen die toch al zichtbaar zijn. Het 'werkelijke' probleem is 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 uitten in tal van hacks en spionageactiviteiten. Fraude wordt al eeuwen gepleegd en deepfakes zijn slechts het volgende middel om vals bewijs in een rechtszaak te introduceren. Zelfs in een wereld zonder deepfakes zouden deze tendensen zich voordoen.

Bovendien is de vrees voor de teloorgang van de waarheid al eeuwen bestaand en speelt die bij de introductie van iedere technologie weer op. De introductie van de drukpers stelde particulieren in staat allerhande pamfletten met opinies en halve waarheden op grote schaal te verspreiden, het internet stelde mensen in staat in bubbels te leven, de virtuele wereld zou de echte wereld op termijn vervangen, etc. Steeds is die vrees slechts gedeeltelijk uitgekomen. Eerder is het zo dat op de introductie van een nieuwe techniek een periode van chaos en ongeregeldheden volgt, waarna er zowel juridische, maatschappelijke als institutionele normen worden ontwikkeld om het gebruik van de techniek in goede banen te leiden. In die zin zijn deepfakes niets nieuws.

Het nieuwe en potentieel gevaarlijke van deepfakes zit in een tweetal zaken, een kwalitatief en de ander kwantitatief. Enerzijds lijken deepfakes zo echt dat zij sneller voor waar zullen worden gehouden. Mensen hebben een zogenoemde 'truth-bias', ze nemen aan dat iets waar is, tenzij er contraindicaties zijn. Dat geldt zeker voor videobeelden. Het tweede, en wellicht nog belangrijkere, verschil is de democratisering van de techniek. De verwachting van alle voor deze studie geïnterviewden was dat de techniek binnen een jaar of twee, drie in handen zou zijn van de gewone burger en grif zou worden gebruikt. Gratis apps zijn nu al beschikbaar en die apps zouden huns inziens alleen maar beter en sneller worden. Het vervaardigen van een zeer realistische deepfake kan dan binnen een handomdraai door vrijwel iedere burger ter wereld. Daarmee zal de hoeveelheid fake-content exponentieel toenemen, zozeer zelfs dat er inderdaad een dreiging kan ontstaan dat echt niet meer van nep te onderscheiden is. Het aandeel nepcontent kan op termijn zelfs de hoeveelheid authentieke content overstijgen. Daarmee wordt het noodzakelijk om binnen veel contexten alle content op authenticiteit te checken, terwijl dit praktisch en financieel gezien vrijwel ondoenlijk zal zijn.

7.6 Handhaafbaarheid

Tot slot wellicht het belangrijkste inzicht wat betreft het juridische kader. Aanpassingen aan het materieel recht en het procesrecht zijn mogelijk en wellicht wenselijk op specifieke punten (zie daarover paragraaf 8.1 en 8.2), maar dat betreft niet het belangrijkste probleem ten aanzien van deepfakes in horizontale verhoudingen en meer in het algemeen van privacyschendingen in horizontale verhoudingen. Op de eerste, tweede en derde plaats staat het handhavingsvraagstuk. Het vervaardigen van pornografisch materiaal van een ander zonder diens toestemming is al verboden; het genereren van kinderporno van een fictief kind is al verboden; het plegen van fraude en misleiding middels een deepfake is al verboden; het aandragen van valselijk bewijsmateriaal in een rechtszaak is al verboden; het aanzetten tot haat of geweld tussen groepen middels een deepfake is al verboden; het zonder toestemming exploiteren van iemands beeld of gelijkenis of creatieve werken is al verboden; het schade berokkenen aan een ander middels een fake-bericht kan al onder het onrechtmatige-daadsregime worden aangepakt; etc.

De juridische inkadering van deepfakes is daarom niet het primaire probleem, het probleem is de handhaving van de bestaande en eventuele aanvullende rechtsregels. Daarbij speelt bij deepfakes, en meer in het algemeen bij internetregulering, een aantal obstakels. Allereerst ontwikkelt de techniek zich snel, zodat regelgeving op specifieke technologie snel achterhaald kan raken. Ook is het lastig de techniek te definiëren voor wetgevingsdoeleinden, omdat een te smalle definitie te veel ruimte laat voor onwenselijke toepassingen of voor technieken die zo worden aangepast dat ze niet onder de gekozen definitie vallen, terwijl een te wijde definitie ook wenselijke technische toepassingen treft en tot rechtsonzekerheid kan leiden. Ten derde zijn er door het grensoverschrijdende karakter van technologieën vaak meerdere juridische regimes, met verschillende regels, op partijen van toepassing en vestigen zij zich doorgaans in de jurisdictie met de laagste regeldruk. Ten vierde is het lastig om de regels uit de ene jurisdictie op te leggen en adequaat af te dwingen ten aanzien van partijen die in andere landen zijn gevestigd. Ten vijfde speelt er vaak een complex web aan partijen die allemaal een deel van het proces onder hun hoede hebben en allemaal een gedeeltelijke verantwoordelijkheid dragen. Ten zesde is het vaak eenvoudig om regels uit een bepaalde jurisdictie te omzeilen, bijvoorbeeld door burgers die middels een VPN-verbinding doen alsof zij in een ander land gelokaliseerd zijn.

Daarnaast zijn twee punten in dit rapport specifiek belicht.

Ten eerste de democratisering van datagedreven technieken, waaronder de deepfaketechnologie. Doordat het aannemelijk is dat alleen al in Nederland binnen een aantal jaar miljoenen gebruikers toegang hebben tot deepfaketechnologie zal het potentiële aantal deepfakes dat op het internet of op afgesloten netwerken verschijnt te groot zijn om alle mogelijke fakeberichten afzonderlijk op authenticiteit te beoordelen. Dat brengt de keuze in het huidige regulerende kader, namelijk om primair in te zetten op *ex post* regulering (de ontwikkeling, het aanbieden en de toegang tot de techniek wordt niet aan banden gelegd, maar het gebruik daarvan voor specifieke doeleinden; dat gebruik wordt pas nadat het materiaal is verspreid en in de openbaarheid is gekomen op

rechtmatigheid getoetst), onder druk. De keuze voor *ex post* regulering en de democratisering van de techniek brengt met zich dat de aandacht en energie van handhavende organisaties vrijwel uitsluitend zal gaan naar de handvol extremere schendingen (vaak relaterend aan de lichamelijke privacy) en dat het overgrote deel van kwalijke, maar niet acuut of bovenmatig problematische deepfakes ongemoeid blijft (het OM treedt slechts in een beperkt aantal zaken op; de AP is vrijwel inactief als het gaat om horizontale privacyschendingen en richt zich op overheden en grote ondernemingen). Dit zorgt er ook voor dat er als vanzelf een normalisering van deze kleine schendingen zal plaatsvinden.

Dit zou kunnen gelden als reden om voor een ander reguleringsmodel te kiezen. Wat *ex ante* regulering echter lastig maakt is dat deepfakes ook voor legitieme doeleinden kunnen worden ingezet. *Ex ante* verboden op het aanbieden of het in bezit hebben van technologie maken ook deze legitieme toepassingen onmogelijk; *ex ante* toetsen op rechtmatigheid van toepassingen is vrijwel ondoenlijk gegeven de hoeveelheid content en zal daarbij ook foutmarges met zich brengen. Dergelijke *ex ante* toetsen roepen natuurlijk ook de vraag op: welke partij beoordeelt of een product of toepassing legitiem is en op basis van welke juridische of morele standaard? Wat een *ex ante* toets op de publicatie van deepfakes bemoeilijkt is dat ze ook kunnen worden vervaardigd met de aanvankelijke toestemming van de betrokkene (een stelletje dat een fake-porno filmpje maakt van henzelf), maar dat de latere verspreiding daarvan (jongen wil wraak nemen op ex-vriendin) niet met toestemming geschiedt. Het is voor een derde (de Autoriteit Persoonsgegevens, een internet intermediair, etc.) vaak niet eenvoudig vast te stellen of beelden en/of de deepfake met toestemming zijn gemaakt en zo ja voor welke doeleinden toestemming is verkregen. Dat vaststellen is vaak een tijdrovende procedure. Daarbij komt dat alhoewel *ex ante* regulering beter te handhaven is, dit nog steeds geen garanties biedt, gezien de territoriale grenzeloosheid van het internet. Een certificeringsregime voor het produceren, aanbieden of downloaden van deepfaketechnologie is om verschillende redenen ingewikkeld. Zo zijn de producenten vaak buiten de EU gevestigd, kunnen apps en diensten worden aangeboden op buitenlandse sites en maakt een verbod volgens de voor deze studie geïnterviewde experts een dienst waarschijnlijk spannender en interessanter en vergt een dergelijk regime een flinke investering in de handhaving.

Ten tweede spelen drie partijen een mogelijke rol bij het toezicht op en de handhaving van de privacynormen in horizontale verhoudingen – de burger zelf, de staat en de tussenpartijen –, maar zijn er belemmeringen ten aanzien van ieder van hen om dit effectief en adequaat te doen.

Iedere burger heeft weliswaar allerhande proces- en klachtrechten, maar het is lang niet altijd duidelijk voor iemand dat zijn data worden of zijn verzameld of dat er een deepfake van hem is verspreid op het internet (bijvoorbeeld op een pornosite of Geenstijl.nl). Zelfs als hij dat wel weet of te weten komt, dan nog is niet altijd duidelijk wie verantwoordelijk kan worden gehouden of aansprakelijk kan worden gesteld. Om achter de identiteit van de dader te komen is vaak de medewerking van internet intermediairs noodzakelijk, maar die willen niet altijd meewerken (zonder last van de rechter) vanwege de privacybelangen van degene die het materiaal heeft geplaatst. Dat betekent dat er vaak twee

rechtszaken nodig zijn, één om de identiteit van de dader te achterhalen en een andere om de dader in rechte aan te spreken. Als het daarbij nog gaat om een verwijderverzoek ten aanzien van het platform of ten aanzien van eventuele kopieën die elders zijn gepubliceerd kan soms een derde, vierde en volgende rechtszaak nodig zijn. Dit vergt tijd, geld en energie die burgers vaak ontberen; de schadevergoeding die wordt geboden als de burger in het gelijk wordt gesteld is doorgaans verwaarloosbaar.

Voor intermediairs is het probleem bij conflicten in horizontale verhoudingen dat lang niet altijd evident is of er een privacyschending heeft plaatsgevonden, of een deepfake een onrechtmatig karakter heeft. Daarnaast gelden er tal van lastige juridische vragen. Worden er persoonsgegevens verwerkt bij een deepfake op basis van beelden van meerdere personen? Is er een legitieme verwerkingsgrondslag voor de deepfake? Geldt er een uitzondering in het kader van de vrijheid van meningsuiting? Vaak staan er voor verschillende burgers verschillende belangen op het spel, zodat de keuze om het verzoek van de één te honoreren tegelijk de keuze behelst om de rechten of belangen van de ander te beperken. Tot slot levert automatische filtersystemen, waardoor bepaalde inhoud wordt geweerd van een site, ook kritiek op, omdat dergelijke systemen altijd zowel over- als onderinclusief zijn.

Voor overheidsinstanties, zoals het Openbaar Ministerie en de Autoriteit Persoonsgegevens, gelden veel van de voorgenoemde problemen evenzeer. Bijvoorbeeld de tijd, moeite en middelen die het kost om allerhande opnames en afbeeldingen op authenticiteit en rechtmatigheid te controleren, de onduidelijkheid die er vaak heerst over toestemming of niet en de diverse complexe juridische vragen die spelen, de vraag of het loont om achter vrij kleine privacyschendingen in horizontale verhoudingen aan te gaan en het probleem dat een keuze voor de bescherming van het recht van de ene burger vaak consequenties heeft voor de rechten van een andere burger.

Voetnoten hoofdstuk 7

- 401 Dobber. T. e.a. (2021), 'Do (Microtargeted) Deepfakes Have Real Effects on Political Attitudes?', *UvA* .
- 402 Ook voor zulke doeleinden is fake kinderporno strafbaar; de politie mag zulk materiaal niet genereren en in netwerken inbrengen. Het is de vraag of behandelaars zulk materiaal naar Nederlands recht zouden mogen gebruiken.
- 403 Zie o.a. Latour, B. (1993). *We have never been modern.* Harvard University Press. Borgmann, A. Technology and the Character of Contemporary Life, University of Chicago Press, 1984.

8. Reguleringsopties

In dit hoofdstuk wordt een breed palet gegeven aan mogelijke reguleringsopties. Daarbij gelden een aantal kanttekeningen:

1. De reguleringsopties zijn geen aanbevelingen, maar opties; de wenselijkheid en haalbaarheid daarvan zullen nader moeten worden onderzocht.
2. Sommige daarvan zijn direct invoerbaar en concreet uitgewerkt, andere betreffen opties voor de lange termijn en vergen structurele wijzigingen in het Nederlandse rechtstelsel. Een specifieke aanpassing in het Wetboek van Strafrecht is bijvoorbeeld relatief eenvoudig in te voeren, andere opties zijn controversieel, hebben grote mogelijk negatieve consequenties en vergen verdergaand onderzoek alvorens tot een eventueel uitgewerkt plan te komen, zoals een verbod of inkadering van het verspreiden van desinformatie als zodanig.
3. De reguleringsopties moeten in onderlinge samenhang worden bezien. Meerdere opties kunnen een onderliggend probleem adresseren; als er wordt gekozen voor de ene optie, kunnen andere dus achterwege blijven. Het geheel of gedeeltelijk verbieden van deepfakes (paragraaf 8.3.1), heeft uiteraard als consequentie dat een deel van de mogelijke aanpassingen in het materieel- en procesrecht achterwege kunnen blijven. De verschillende opties als besproken in paragraaf 8.2 zien allen op hetzelfde onderliggende probleem; zij kunnen deels als elkaar aanvullend worden beschouwd, maar ze allemaal invoeren zal hoe dan ook niet voor de hand liggen. Reguleringsoptie 1 en 5 zien in wezen op hetzelfde onderliggende probleem, namelijk het feit dat de verwerking van persoonsgegevens over anderen in de privésfeer thans ongereguleerd is. Ook hier kunnen beide opties worden ingevoerd en als elkaar aanvullend worden beschouwd, maar kan ook met een van de twee worden volstaan, mocht de wetgever dit punt willen adresseren.
4. Veel van de beschreven problemen en mogelijke oplossingen hangen samen met meer algemene, maatschappelijke tendensen. Het probleem van desinformatie is niet eigen aan deepfakes, alhoewel zij daar wel aan kunnen bijdragen; het probleem van het besmeuren van de eer en goede naam van politici is niet uniek voor deepfakes, maar kan daardoor wel een extra stimulans krijgen; het gevaar voor beïnvloeding van verkiezingen staat los van deepfakes reeds op de politieke agenda, maar wordt door deepfaketechnologie nog eens extra op scherp gezet. Soms ligt het voor de hand om specifieke regels voor deepfakes aan te nemen; vaak is het echter raadzamer om het onderliggende probleem als zodanig te adresseren, waardoor ook bepaalde toepassingen van deepfakes worden gereguleerd, als onderdeel van een bredere rechtsregel.

De reguleringsopties geven dus een palet aan mogelijkheden voor de Nederlandse wet- en regelgever. Ze zijn gerangschikt aan de hand van de inzichten die zijn verkregen in hoofdstuk 3 (het materieelrecht), hoofdstuk 4 (het procesrecht) en hoofdstuk 5 (handhaving en toezicht). Paragraaf 8.1 geeft de diverse mogelijkheden om het materiele recht aan te passen, paragraaf 8.2 doet dat ten aanzien van het procesrecht en paragraaf 8.3 bekijkt hoe het toezicht op en de handhaving van de vigerende regels ten aanzien van deepfakes kunnen worden versterkt.

8.1 Materieel recht

Voor het materieel recht geldt dat een klein aantal aanpassingen mogelijk is. Meer in het algemeen is het zo dat veel van de huidige juridische normen in onder meer het strafrecht, het privacy- en gegevensbeschermingsrecht en het intellectueel eigendomsrecht open en algemeen van aard zijn en dat er discussie is en kan ontstaan over de precieze interpretatie en toepassing van deze regels op het fenomeen deepfake. Hierin kunnen verschillende partijen een proactieve rol spelen. De overheid kan nieuwe, specifieke normen stellen en instanties als de Autoriteit Persoonsgegevens en het Openbaar Ministerie kunnen aanwijzingen en richtsnoeren uitvaardigen. Als dat niet gebeurt zal deze invulling en uitwerking uiteindelijk door de rechter gebeuren. Niets doen is wat dat betreft derhalve ook een optie, al duurt het dan langer voordat er rechtszekerheid op een aantal punten wordt verkregen.

Dan is er nog een aantal punten waar nieuwe regels voor kunnen worden aangenomen. Achtereenvolgens zullen deze reguleringsopties binnen het strafrecht, het privacy- en gegevensbeschermingsrecht, het vrijheid van meningsuitingsregime en het civiel recht worden besproken.

8.1.1 Strafrecht

Het materiële strafrecht is 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. Verreweg de meeste delicten zijn immers voldoende technologie-neutraal geformuleerd voor wat betreft de vorm waarin deze kunnen worden gepleegd.

Voor seksfilmpjes die met deepfaketechnologie zijn gemanipuleerd zodat het lijkt alsof een bestaande persoon daarin figureert – een veelvoorkomende en mogelijk voor slachtoffers de meest ingrijpende vorm van deepfakes – is het belangrijk te constateren dat daarop de strafbaarstelling van wraakporno (art. 139h lid 2 onder b Sr) toepasbaar is, althans voor zover zulke filmpjes openbaar worden gemaakt en dat nadelige gevolgen voor de afgebeelde persoon kan hebben. Aan die laatste voorwaarde zal meestal zijn voldaan, en aangezien openbaarmaking in deze bepaling ruim wordt uitgelegd als het delen met een of meer anderen, zal artikel 139h lid 2 onder b Sr goed kunnen worden ingezet om seksuele deepfakes tegen te gaan. Daarnaast levert het verspreiden van zulke filmpjes mogelijk ook smaadschrift op, en wanneer de video ook identificerende persoonsgegevens (zoals de naam van de gedeepfakete persoon) bevat, zal bovendien artikel 231b Sr (identiteitsfraude) van toepassing zijn.

Wanneer deepfake-seksvideo's echter niet worden verspreid, maar puur voor eigen gebruik worden gemaakt en bekeken, valt dit niet onder een strafbepaling. Het is een rechtspolitieke vraag of het voor eigen gebruik maken van zulke deepfakes van iemand zonder diens toestemming strafbaar zou moeten worden gesteld. Daarover zou wellicht een debat kunnen worden gevoerd, dat eventueel ook breder zou kunnen worden betrokken op alle vormen van deepfakes die in een bepaald opzicht onterend kunnen worden geacht voor de gedeepfakete persoon en daarmee wellicht als "intrinsiek" strafwaardig zouden kunnen worden beschouwd. Aangezien de wetgever tot nu toe alleen het bezit van kinderpornografie en dierenpornografie als "intrinsiek" strafwaardige gegevens heeft

strafbaar gesteld, en niet andere vormen van inhoud die in het maatschappelijk verkeer vaak als moreel verwerpelijk worden beschouwd (denk aan onthoofdingsvideo's of *Mein Kampf*), lijkt wel terughoudendheid gepast om het maken of bezitten van "onterende" deepfakes als zodanig strafbaar te stellen.

Reguleringsoptie 1:

Voer een debat over de vraag of het maken of bezitten van "intrinsiek" strafwaardige (moreel verwerpelijke) deepfakes strafbaar zou moeten worden gesteld.

De enige mogelijke lacune in de wetgeving die in dit onderzoek is geconstateerd, betreft het gat 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 als nadeel kan ontstaan. 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. Voor deepfakes levert dit niet per se een lacune in de rechtsbescherming op, aangezien zoals gezegd de meeste strafbepalingen door hun technologie-neutrale formulering van toepassing zijn. Mocht de wetgever het echter 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'. Hierdoor zou immers een algemene strafbaarstelling ontstaan van misbruik van iemands gelaat of stem als daaruit enig nadeel kan ontstaan.

Reguleringsoptie 2:

Overweeg artikel 231b Sr aan te passen door de clausule 'niet zijnde biometrische gegevens' te schrappen of aan te passen.

8.1.2 Privacy- en gegevensbeschermingsrecht

Binnen het privacy- en gegevensbeschermingsrecht zouden drie additionele bepalingen kunnen worden overwogen ten aanzien van deepfakes, waarmee ook een breder vraagstuk wordt geadresseerd, namelijk de introductie van een post-mortem privacy recht, een begrenzing aan het creëren van nieuwe personen en de inperking van de huishoudelijke exceptie.

Het vraagstuk omtrent post-mortem privacy is al decennia oud, maar staat zeker de laatste jaren weer vol in de discussie.[404] Naarmate er meer data over personen beschikbaar zijn is ook de waarde van die data na diens overlijden steeds groter. Nu is het zo dat als een persoon overlijdt, deze gegevens in principe niet onder het beschermingsregime van de AVG vallen, tenzij ze bijvoorbeeld indirect iets zeggen over nog levende personen (in welk geval ze als persoonsgegevens van de nog levende persoon kunnen worden gezien). Veel burgers willen niet dat hun gegevens vrijkomen na hun dood; zij willen hun e-mails bijvoorbeeld vernietigd zien als zij zijn overleden. Anderzijds willen nabestaanden vaak juist toegang tot die gegevens, zeker als

degene bijvoorbeeld zelfmoord heeft gepleegd. Ook is de vraag wie de data van een overleden persoon mag exploiteren. Nu zijn bedrijven vaak in het bezit van de data en blijven die verwerken en gebruiken, ook na de dood van een persoon (zij het vaak in geaggregeerde datasets); medische organisaties die lichaamsmateriaal van een overleden persoon hebben willen daar graag wetenschappelijk onderzoek mee blijven uitvoeren; nabestaanden kunnen de wens hebben de data van en over een overleden persoon exploiteren; etc.

Deepfakes trekken deze discussie naar een nieuwe dimensie, zowel op moreel als op commercieel vlak. In principe staat er privacyrechtelijk niets aan in de weg om een overleden persoon, of diegene dat nu wilde of niet, weer tot leven te wekken, bijvoorbeeld door hem te laten figureren op zijn eigen begrafenis of op dagelijkse basis te laten communiceren met de nabestaanden, ook al kan dat indruisen tegen de specifieke wensen van die persoon. Op andere punten biedt het recht wel additionele morele bescherming aan de rechten/belangen van doden. Zo is de omgang met het lichaam van een overledene aan strikte regels en vereisten verbonden, niet alleen vanwege de hygiëne, maar ook vanwege morele fatsoensnormen. Dergelijke regels betreffen het fysieke lichaam, deepfakes geven een realistische weergave van dat lichaam en sterker nog, kunnen de suggestie van iemands psyche en gedachtewereld geven. In die zin zijn deepfakes van overleden personen potentieel wellicht even inbreuk makend als disresepectvolle handelingen met het lichaam van een overleden persoon.

Daarnaast zijn er tal van toepassingen waarover een maatschappelijk discussie en politiek debat zou kunnen worden gevoerd. Is het bijvoorbeeld wenselijk en toegestaan om overleden historische figuren op scholen les te laten geven? Is het wenselijk en toegestaan om overleden kunstenaars een rondleiding te laten geven in een museum? Is het wenselijk en toegestaan om overleden acteurs in films te laten figureren? Is het wenselijk en toegestaan om een overleden persoon in een pornofilm te laten spelen? Is het wenselijk en toegestaan om overleden artiesten nog concerten te laten geven? Etc.

Reguleringsoptie 3:

Bekijk in hoeverre het creëren van wet- of regelgeving aangaande post-mortem privacy wenselijk is

Het creëren van niet bestaande personen roept tal van ethische dillema's op. Allereerst is duidelijk dat hierdoor wordt bijgedragen aan de vervaging van wat echt is en wat nep. Dat kan in concrete gevallen alsnog wenselijk zijn, bijvoorbeeld als de politie door middel van een fake persoon kan infiltreren in een crimineel netwerk, als door middel van fakekinderporno pederasters kunnen worden opgespoord of door middel van fake klanten vrouwenhandelaren in kaart kunnen worden gebracht. Toch is zelfs in deze gevallen terughoudendheid geboden.

Zowel voor deze toepassingen als voor andere toepassingen ligt het voor de hand om meer duidelijkheid te geven omtrent wat wel en wat niet is toegestaan in termen het creëren en inzetten van fictieve, maar zeer realistische personages, niet alleen door de politie, maar ook binnen de entertainmentindustrie, binnen de porno-industrie of voor medische toepassingen. Zo zijn er therapieën voor de behandeling van

pedofielen door middel van het tonen van nep-kinderporno, maar is dat wenselijk? Nu al geldt er een verbod voor datingsites om mannen middels nepprofielen het idee te geven dat zij spreken met een dame van vlees en bloed, terwijl dat niet het geval is. Dat is op zijn plaats in het kader van misleiding binnen de contractuele relatie tussen klant en datingsite, maar hoe zit dat met deepfake porno? Ook zijn er mogelijke morele grenzen ten aanzien van wat een fake-persoon mag doen of laten.

Reguleringsoptie 4:

Bekijk in hoeverre het creëren van wet- en regelgeving omtrent het gebruik van volledig door AI gegenereerde personen wenselijk is

Tot slot is er binnen het privacy- en gegevensbeschermingsrecht een aanpassing mogelijk ten aanzien van de huishoudelijke exceptie. De exceptie, die uit de Richtlijn 1995 stamt en die bepaalt dat persoonsgegevens die worden verwerkt in de privésfeer in principe niet aan de regels uit het gegevensbeschermingsrecht hoeven te voldoen, stond reeds ter discussie toen de AVG werd vervaardigd, maar is toch vrijwel ongewijzigd gelaten. In hoeverre Nederland hier zelfstandig alternatieve regels omtrent kan aannemen – in hoeverre de AVG hiertoe aan nationale lidstaten ruimte laat – is niet met zekerheid te zeggen. Als alternatief zou er kunnen worden gekozen om op Europees niveau te pleiten voor nadere regelgeving op dit punt. Door aanpassing en begrenzing van de huishoudelijke exceptie zouden twee punten worden aangepakt: een normatieve kwestie en een handhavingskwestie.

Allereerst kan de verwerking van persoonsgegevens over anderen binnen de privésfeer ook problematische vormen kan aannemen. Toen de Richtlijn 1995 werd aangenomen werd nog primair gedacht aan het bijhouden van een adressenboek. Daarbij worden weliswaar de persoonsgegevens van derden verwerkt, maar dat betreft slechts de naam, het adres en het telefoonnummer. Het bijhouden van dergelijke gegevens is bovendien sociaal geaccepteerd en doorgaans gewenst door de derden in kwestie. Sindsdien zijn er echter veel nieuwe vormen van gegevensverwerking in handen van burgers gekomen, waarvan deepfaketechnologie wellicht het voorlopige hoogtepunt vormt. Stel een ex-partner bewaart op zijn computer privéfoto's van zijn ex-vriendin, waarmee hij vervolgens een deepfake maakt waarin zij allerhande perverse seksuele handelingen verricht. Hij vertelt daar vervolgens over aan zijn vrienden, die dat ook aan haar communiceren. Dit is maar een van de vele mogelijke voorbeelden van deepfake-toepassingen die wel kwalijk en schadelijk kunnen zijn, maar momenteel niet als zodanig onder de AVG geadresseerd kunnen worden.

Ten tweede zou met de begrenzing van de huishoudelijke exceptie ook een handhavingsprobleem worden geadresseerd. Het denken in sferen – bijvoorbeeld de privésfeer als onderscheiden van de publieke ruimte – is voor de hand liggend in een wereld waarin die sferen ook in reële zin van elkaar gescheiden zijn en waarbij de grenzen tussen die sferen tamelijk duidelijk zijn. Dat is echter al lang niet meer zo. De mobiele telefoon vormt vaak de poort tot de meest intieme privégegevens, terwijl die ook in de publieke ruimte wordt meegenomen. De privésfeer is steeds minder een afgesloten ruimte en steeds meer een poreuze omgeving, waarin diverse 'slimme' apparaten permanent in verbinding staan met het internet en

gegevens over wat zich in de privésfeer voltrekt doorspelen aan derde partijen. Ook het gemak waarmee gegevens vanuit de privésfeer naar een miljoenenpubliek kunnen worden gebracht is wezenlijk anders; nu is een klik op de knop voldoende om duizenden foto's of video's die op een privécomputer staan online te verspreiden via sociale media of digitale platformen. De huishoudelijke exceptie brengt het volgende probleem met zich. Het vervaardigen van compromitterend materiaal, in het hierboven gegeven voorbeeld van een deepfake van een ex-geliefde, en het in bezit hebben daarvan, valt niet onder de AVG. Als het materiaal eenmaal op het internet staat of onder grote groepen vrienden is verspreid wel, maar dan is het al te laat. De schade heeft zich dan al voltrokken; de ervaring leert dat filmpjes van sensationele aard binnen enkele uren duizenden kijkers kunnen trekken. De schade voor het slachtoffer is dan al onherroepelijk.

Daarom zou kunnen worden overwogen om de huishoudelijke exceptie te beperken. Wel moet daarbij worden bedacht dat alhoewel dit het probleem van handhaving verkleint omdat kwaadaardige content al in een vroegtijdig stadium kan worden tegengegaan en de verspreiding daarvan onder een groot publiek, met alle problemen van dien - zoals de eenvoud waarmee kopieën kunnen worden gemaakt en vervolgens weer verder worden verspreid, ook als het oorspronkelijke materiaal offline is gehaald - het ook een nieuw handhavingsprobleem oproept. Hoe kan ervoor worden gezorgd dat alle normen ook in de privésfeer worden gehandhaafd? Als de overheid daarvoor zorg tracht te dragen, is dan de kuur niet erger dan de kwaal?

Reguleringsoptie 5:

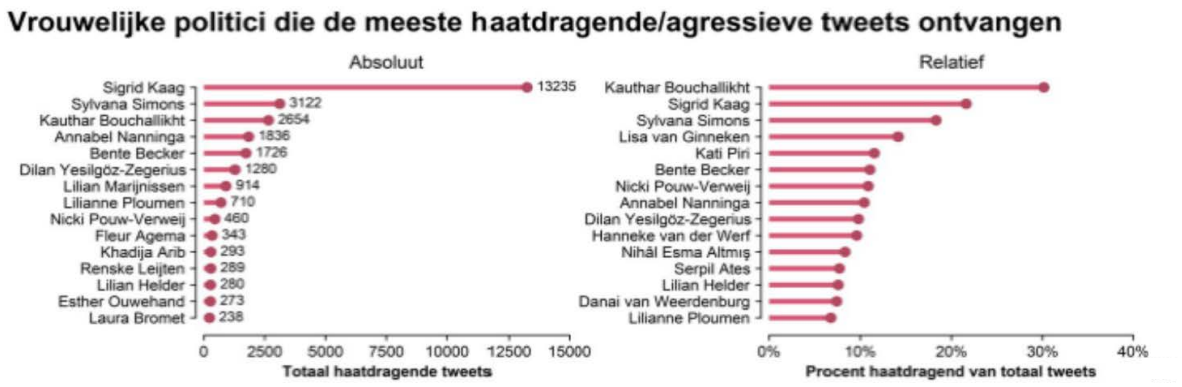
Bekijk in hoeverre de herziening van de huishoudelijke exceptie binnen het gegevensbeschermingsrecht wenselijk is

8.1.3 Vrijheid van meningsuiting en recht op reputatie

Binnen het kader van de vrijheid van meningsuiting verdient een drietal vraagstukken aandacht, namelijk de positie van publieke figuren, de verspreiding van foutieve informatie en de impact van fake-nieuws op verkiezingen.

Deepfakes ingezet door burgers zijn grosso modo te verdelen in twee groepen. Deepfakes die zich richten op henzelf of mensen in hun omgeving, zoals hun partner, kinderen, ouders, buurtgenoten en vrienden, en deepfakes die zich richten op publieke figuren, zoals acteurs, zangers, politici, ambtenaren, royalty en journalisten. Over het algemeen zullen burgers sneller toegang hebben tot privégegevens van personen uit de eerste groep terwijl zij, in het algemeen, de beelden over personen uit de tweede groep vaker uit openbare bronnen, zoals het internet zullen halen. Over het algemeen kan van personen uit de eerste groep gemakkelijker toestemming worden gevraagd, terwijl dit voor personen uit de tweede groep doorgaans lastiger is. Dat betekent dat voor de tweede groep meestal geen verwerkingsgrondslag voor bijzondere persoonsgegevens – aangaande bijvoorbeeld seksuele informatie, politieke overtuiging of medische condities – voor handen zal zijn en voor gewone persoonsgegevens altijd moet worden gekeken naar zowel de belangen van de burger die de deepfake maakt als naar die van de bekende persoon in kwestie.

405



Tegelijkertijd zijn er binnen het regime omtrent de vrijheid van meningsuiting en het recht op reputatie bijzondere regels voor publieke figuren. Alhoewel zij nog steeds een beroep kunnen doen op de bescherming van hun eer en goede naam en zelfs voor wat betreft hun gedragingen in de publieke ruimte een beroep op privacy kunnen doen, heeft het Europees Hof voor de Rechten van de Mens ook bepaald dat publieke figuren een grotere inmenging in hun privéleven moeten dulden dan gewone burgers, dat zij moeten accepteren dat zij zullen worden bespot en beschimpt en dat er over hen en hun gedragingen gepeperde en soms vergaande meningen worden verkondigd. Hoe de verhouding van de vrijheid van meningsuiting van burgers en het recht op bescherming van de eer en goede naam van publieke personen dient te worden beoordeeld, daar heeft het Europees Hof slechts een aantal zeer algemene principes voor neergelegd. Het echte oordeel zal van zaak tot zaak geschieden, rekening houdend met de concrete omstandigheden van het geval.

Ook dit sluit aan bij een bredere problematiek. De vaagheid brengt met zich dat publieke figuren van tevoren weinig tot geen rechtszekerheid hebben omtrent de bescherming van hun persoonlijke levenssfeer en dat aan burgers geen duidelijke regels worden gesteld voor wat betreft hun uitlatingen over bekende personen. Dit draagt eraan bij dat er slechts zelden wordt opgetreden tegen online uitingen, bijvoorbeeld als er een filmpje van Sylvana Simons wordt vervaardigd waarin zij lijkt te worden verhangen door de Ku Klux Klan. Het overige wordt vaak oogluikend toegestaan en noch het OM noch de bekende personen zelf ondernemen actie tegen dubieuze uitingen. Dit heeft een normalisering van extreme uitingen over publieke personen tot gevolg.

Vaak hebben ook deze uitingen weer een misogyne ondertoon. Koploper in Nederland is hierbij Sigrid Kaag, onder meer omdat zij getrouwd is met een Palestijnse man, gevolgd door Sylvana Simons.[406] Voor vrouwelijke politici blijkt dat de voortdurende inbreuken op hun eer, goede naam en persoonlijke levenssfeer een belangrijke overweging te zijn bij de keuze om in of uit de politiek te gaan. Het feit dat deze beledigingen plaatsvinden is uiteraard niet nieuw, wel de massaliteit daarvan en het ophitsende karakter van dergelijke opmerkingen op discussiefora en sociale media.[407] Bij mannen gaat het vaker om doodsbedreigingen, waarvan zij vaker niet dan wel aangifte doen, die vaker niet dan wel tot vervolging leiden,[408] die vaker niet dan wel tot een veroordeling leiden. Voor zowel mannen als vrouwen geldt steeds meer dat de

online cultuur een reden is om geen publieke functie te bekleden of niet op de voorgrond te treden. Dat kan op termijn schadelijke gevolgen hebben voor de democratie en het openbaar bestuur als zodanig.

Deepfakes zullen deze bestaande problematiek verergeren. Veel deepfakes gaan bijvoorbeeld over gezagsdragers. Politici die dronken een speech lijken te geven, bestuurders die rare woorden in de mond worden gelegd, politieagenten die geweld lijken te gebruiken, ambtenaren die de regels lijken te overtreden waarop zij zelf zouden moeten toezien, een staatshoofd dat een nep kerstoespraak geeft, etc. Dit soort fake-berichten sluit aan bij de sentimenten die binnen een bepaalde groep in de samenleving toch al leven en zullen hun wereldbeeld versterken. Dit kan radicalisering en polarisering in de hand werken.

Daarom zou de Nederlandse wetgever kunnen overwegen om nadere regels te stellen ten aanzien van de bescherming van de eer en goede naam van publieke personen. Hierbij zou duidelijker kunnen worden aangegeven welke grenzen er zijn ten aanzien van burgerlijke uitingen over publieke personen, zodat beide zijden (de burger die een uiting doet en de publieke persoon) beter weten waar zij aan toe zijn. De burger kan dan op basis van het kader zijn gedrag en uitlatingen aanpakken, publieke figuren en eventueel handhavende organisaties kunnen van tevoren beter inschatten of een rechtsgang de moeite loont. Een dergelijk initiatief zou aansluiten bij een groeiend besef dat het van essentieel belang is om het gezag van volksvertegenwoordigers, gezagsdragers en ambtenaren beter te beschermen tegen publieke uitingen, niet alleen om hun persoonlijke belangen beter te waarborgen, maar ook om het tanende vertrouwen in de overheid tegen te gaan.[409] Daarbij moet uiteraard wel het belang van een open en vrij debat centraal staan.

Reguleringsoptie 6:

> Bekijk inhoeverre het aannemen van wet- of regelgeving waarin nadere regels worden gesteld met betrekking tot uitingen over publiek figuren wenselijk is

Meer in het algemeen is een vrees ten aanzien van deepfakes dat zij het post-truth tijdperk naar een hoger-niveau zullen tillen. Deepfakes zijn per definitie niet echt en vergroten daarmee de verwarring die nu al ontstaat omtrent de waarachtigheid en juistheid van content. Het doen van onware uitingen is momenteel niet an sich gereguleerd. Een onware, onjuiste of misleidende uiting kan onder het huidige regime wel worden geadresseerd, maar slechts als er schade is ontstaan, bijvoorbeeld aan persoonlijke belangen (onder het onrechtmatige-daadsregime) of aan bepaalde maatschappelijke belangen, zoals haat tegen bepaalde groepen wordt aangewakkerd (onder het strafrecht).

Dit brengt twee zaken met zich. Ten eerste dat er een ingewikkeld vraagstuk ontstaat ten aanzien van de causale relatie tussen een onware, onjuiste of misleidende uiting en de (te verwachten) schade die dat veroorzaakt. Ten tweede dat onware, onjuiste of misleidende uitingen niet als zodanig zijn gereguleerd, terwijl zij als zodanig wel problematisch kunnen zijn en in ieder geval bijdragen aan een wereld waarin feiten en meningen, waarheden, leugens en halve-waarheden steeds meer door elkaar lopen. Daarom zou hier regulering op kunnen worden aangenomen, bijvoorbeeld door duidelijk

onware, onjuiste of misleidende uitingen als zodanig aan banden te leggen.

Op zich pleit er weinig tegen het inkaderen van dergelijke uitingen, behalve dat er een scheidsrechter moet zijn van wat waar is en wat niet. Het is niet wenselijk dat de overheid hier een actieve rol in speelt; het gevaar van een overheid die een monopolie claimt op de waarheid is volgens velen groter dan het gevaar van onware, onjuiste of misleidende informatie. De ervaringen die er op EU-niveau zijn met het tegengaan van fake-nieuws laat ook zien hoe complex dergelijke inschattingen zijn en hoe snel er foute keuzes kunnen worden gemaakt ten aanzien van wat waar is en wat niet, wat vervolgens leidt tot grote ophef en een nog groter wantrouwen jegens de overheid.

Toch hoeft de overheid hier geen actieve rol in te spelen. Het kan ook aan de rechter worden gelaten om hier een keuze in te maken. De rechter hoeft niet te treden in complexe vragen ten aanzien van wat waar is en wat niet, maar kan wel evidente onwaarheden zoals die tijdens de Corona-crisis voordeden – met een vaccinatie wordt een chip in je lichaam geplaatst, om zo controle over onze lichaam en geest te krijgen, om maar een voorbeeld te geven – aanpakken. Dergelijke uitingen zijn nu niet of nauwelijks juridisch te adresseren, terwijl hun impact op de vaccinatiebereidheid significant is. Wederom speelt dat dit probleem niet uniek is aan deepfakes, maar dat deepfakes wel aan dit fenomeen bijdragen

.

Reguleringsoptie 7:

Bekijk in hoeverre het wenselijk is om wet- of regelgeving te creëren ten aanzien van evident onware, onjuiste of misleidende uitingen

Tot slot speelt het vraagstuk van misleiding specifiek ten aanzien van verkiezingen. Hieromtrent zijn in de Kamer reeds meerdere debatten gevoerd, er zijn verschillende rapporten verschenen omtrent dit onderwerp en er is een breed palet aan maatregelen voorgesteld om het probleem tegen te gaan, zoals het actief doen aan fact-checking, het tegenspreken van valsheden, het voorlichten van mensen over het probleem van misleidende informatie, het inkaderen van politieke advertenties en het opleggen van nadere zorgplichten aan sociale media en internet tussenpersonen.[410]

Toch gaat deze regelgeving niet zo ver als in sommige voor dit rapport bestudeerde landen. Zo zijn er diverse staten binnen de Verenigde Staten die nadere regulering hebben aangenomen ten aanzien van het verspreiden van misinformatie ten tijde van verkiezingen. Die wetten zijn vaak direct of indirect gekoppeld aan deepfakes, maar kennen ook een breder toepassingsbereik. In de besproken voorstellen wordt een verscheidenheid aan toepassingen van deepfakes ingekaderd en wel op verschillende wijzen. Dit zou in Nederland aanleiding kunnen zijn om soortgelijke wetgeving te introduceren. Zo kunnen er civielrechtelijke acties mogelijk worden gemaakt voor politici of partijen die tijdens verkiezingscampagnes slachtoffer zijn geworden van deepfakes, maar kan er ook strafrechtelijk worden opgetreden tegen de beïnvloeding van verkiezingen als zodanig.

Bij deze beïnvloeding wordt vaak gewezen op het gevaar van inmenging door buitenlandse mogendheden en dan met name Rusland. Dit gevaar is er zeker en moet ook voor de Nederlandse democratie niet worden onderschat. Het is zo eenvoudig om middels deepfakes valse geruchten te verspreiden dat er geen wezenlijke drempel is voor kwaadwillende staten en groepen om zich te mengen in buitenlandse verkiezingen. In de voor deze studie gehouden interviews wezen experts overigens nog op een ander gevaar, namelijk dat Rusland, China, Iran en soortgelijke staten hun pijlen niet zozeer zouden richten op westerse democratieën, maar op tweede en derdewereldlanden (the Global South). Dit zou kunnen geschieden voor allerhande doeleinden, zoals het beïnvloeden van concrete besluitvorming (een Russisch staatsbedrijf dat een contract krijgt in plaats van een Amerikaans bedrijf), het beïnvloeden van verkiezingen aldaar (bijvoorbeeld om een Xi-angehaucht regime aan de macht te krijgen) of het beïnvloeden van besluiten op internationaal niveau (bijvoorbeeld door hen te bewegen voor het opheffen van de sancties tegen Iran te stemmen). Daarom pleitten deze experts ervoor dat Westerse democratieën niet alleen zorgen dat hun eigen verkiezingen en politieke besluitvorming niet wordt beïnvloed, maar ook dat andere landen daar vrij van blijven. Zo niet, dan kunnen de effecten op internationaal en strategisch niveau significant zijn.

Evenzo kunnen verkiezingen echter worden beïnvloed door burgers en groeperingen uit het eigen land. Burgers verzinnen nu reeds allerhande theorieën en complotten, geven politici de schuld van zaken waar zij geen verantwoordelijkheid voor dragen en proberen hen welgevallige politici in een goed daglicht te zetten. Strafrechtelijke en civielrechtelijke bepalingen, hebben slechts een beperkt effect op het gevaar van buitenlandse inmenging. Daar liggen andere instrumenten voor de hand, zoals diplomatie of eventuele digitale tegenacties. Voor de beïnvloeding van verkiezingen en politieke besluitvorming, al dan niet middels een deepfake, door burgers en groeperingen van het land zelf liggen deze juridische instrumenten juist wel voor de hand.

Reguleringsoptie 8:

Bekijk in hoeverre het wenselijk is om wet- of regelgeving te ontwikkelen ten aanzien van het beïnvloeden van verkiezingen of politieke besluitvorming middels het creëren of verspreiden van misinformatie

8.2 Procesrecht

Binnen het procesrecht kunnen deepfakes een grote rol gaan spelen. Uit de voor deze studie gehouden interviews blijkt dat zowel in de opsporings- en onderzoeksfase als in de rechtszaal doorgaans wordt aangenomen dat bewijs en meer in het bijzonder films authentiek zijn, tenzij er contra-indicaties zijn of een verdachte stelt dat het bewijs niet correspondeert met de werkelijkheid. Als deepfaketechnologie verder democratiseert en de voorspelling dat op termijn zo´n 90% van de digitale content gemanipuleerd zal zijn uitkomt, dan zal deze vuistregel wellicht moeten worden herzien. De politie is zich momenteel aan het oriënteren op procedures en kaders voor het controleren van bewijsstukken in het deepfake-tijdperk. Binnen de rechtszaal zijn geen initiatieven bekend om met deze nieuwe realiteit om te gaan.

De aanname dat er van de authenticiteit van bewijs mag worden uitgegaan, tenzij er contra-indicaties zijn, betekent dat het met name aan de verdachte of binnen een civiele procedure aan de wederpartij is om te stellen en eventueel te bewijzen dat bewijs gemanipuleerd of gefabriceerd is. De vraag is of dat wenselijk is, enerzijds omdat het een privatisering van een algemeen probleem met zich brengt en anderzijds omdat een burger niet altijd in staat zal zijn om de waarachtigheid en juistheid van bewijs te betwisten. Daarbij kan niet alleen worden gedacht aan verdachten die bij verstek worden veroordeeld, maar bijvoorbeeld ook aan personen met mentale stoornissen. Daarnaast kan het kostbaar zijn om middels technische expertise aan te tonen dat bewijs al dan niet authentiek is, wat met name voor zwakkere partijen, bijvoorbeeld in economisch achtergestelde posities, een bezwaar kan zijn. Er zou derhalve kunnen worden overwogen om verplichtingen of zorgplichten op te leggen bij andere partijen dan de burger, teneinde er zorg voor te dragen dat het algemene belang, namelijk dat rechtszaken worden gewezen op basis van authentiek bewijs, niet slechts afhangt van de assertieve burger.

Daarom kan het wenselijk zijn om na te denken over eventuele nadere juridische bepalingen. Daarbij liggen drie mogelijk partijen voor de hand: de advocaat, politie en justitie en de rechter. Ook is het mogelijk om een zware straf op te leggen aan burgers of procespartijen die willens en wetens frauderen met bewijs.

- De advocaat heeft een professionele zorgplicht om te zorgen dat er slechts waarachtig bewijs wordt aangedragen. Tegelijkertijd dient hij de belangen van zijn cliënt naar voren te brengen en diens versie van de waarheid over het voetlicht te brengen. Die versie van de waarheid is niet altijd de versie die de rechtbank uiteindelijk zal hanteren; dat betekent dat de vraag in hoeverre de advocaat zaken naar voren mag brengen die later onwaar blijken te zijn niet eenduidig te beantwoorden is; dat geldt ook voor de vraag in hoeverre hij bewijs, bijvoorbeeld dat hij van zijn cliënt heeft gekregen, op authenticiteit dient te checken. Een dergelijke plicht brengt bovendien een zeer grote verantwoordelijkheid en investering in tijd en eventuele middelen met zich om materiaal op waarachtigheid te controleren. Tot slot moet worden bedacht dat in niet alle processen een burger vertegenwoordigd wordt door een advocaat. Toch is juist de optie om nadere plichten aan advocaten op te leggen meermaals genoemd tijdens de interviews, met name omdat de advocaat als professionele partij wordt geacht een centrale rol te hebben in de goede procesgang. Daarom is een optie om nadere juridische verplichtingen te stellen, een andere om te kijken hoe de huidige plichten beter en eventueel zwaarder kunnen worden gehandhaafd en tot slot lijkt zelfregulering, bijvoorbeeld door de Orde van Advocaten die een richtlijn opstelt aangaande het verifiëren van in processen ingebracht bewijs wellicht de meest voor de hand liggende optie.
- Daarnaast zou binnen het strafprocesrecht kunnen worden gekeken naar een nadere rol voor politie en justitie. Dit is wenselijk omdat alleen al een justitieel onderzoek of vervolging voor strafrechtelijke delicten een grote impact kan hebben op de verdachte, zijn persoonlijke leven en maatschappelijke status. De politie is reeds bezig standaarden te ontwikkelen voor het verifiëren van bewijs. Dergelijke zelfreguleringsinitiatieven kunnen worden afgewacht en ook kan er worden ingezet op meer bewustwording binnen de diverse organisaties en organisatieonderdelen

ten aanzien van het bestaan en het gevaar van deepfakes. Ook zou kunnen worden nagedacht over een eventuele plicht waarbij politie slechts een persoon als verdachte mag aanmerken en het OM slechts bewijs mag aandragen in een rechtszaak nadat het op authenticiteit is gecontroleerd, bijvoorbeeld door een combinatie van deepfakedetectietechnieken en menselijke evaluatie door experts. Hiertoe zouden financiële middelen moeten worden vrijgemaakt.

- Hierop aansluitend zou kunnen worden bekeken in hoeverre er een plicht voor rechters kan worden opgenomen om bewijs te verifiëren op authenticiteit alvorens dat toe te laten in een rechtszaak. Daarbij is van belang dat niet alleen de voorspelling is dat meer dan 90% van alle online content over een aantal jaar geheel of gedeeltelijk gefabriceerd of gemanipuleerd zal zijn, maar ook dat deepfake niet slechts om video of beeld gaat, maar ook om audio, tekst en bijvoorbeeld satellietsignalen. Elk type gegeven en elke drager kan dus zijn gemanipuleerd. Het zou dus in principe moeten gaan om alle bewijsstukken.
- Tot slot is geopperd om een zwaardere sanctie te introduceren voor burgers die ofwel in een civiele zaak ofwel als verdachte in een strafzaak gemanipuleerd materiaal aanleveren. Zo zou een strafrechtelijk speciaal verbod op het aanleveren van deepfake-bewijs kunnen worden geïntroduceerd. Een dergelijke bepaling zou een afschrikwekkende werking kunnen hebben en als het inderdaad een keer tot een veroordeling komt, zal dat een signaalfunctie hebben. Of deze strafrechtelijke optie proportioneel en wenselijk is zal moeten worden beoordeeld. Daarbij moet worden opgemerkt dat er momenteel reeds een juridische verplichting is om slechts waar en authentiek bewijs aan te dragen in rechtszaken en dat advocaten die willens en wetens sjoemelen met bewijs daar zowel tuchtrechtelijk als anderszins op kunnen worden aangesproken. Toch rees uit de voor deze studie gehouden interviews het beeld op waarin deze bepalingen in de praktijk slechts van beperkt belang zijn. Slechts sporadisch worden advocaten die de fout in gaan tuchtrechtelijk aangepakt en bij de constatering dat in civiele procedures een van de partijen vals bewijs heeft aangedragen volgt vaak uitsluitend dat die partij de zaak verliest, maar worden geen verdere boetes uitgedeeld of sancties opgelegd.

Om dit proces te vergemakkelijken zijn twee mogelijkheden naar voren gekomen, namelijk ten eerste om een instituut op te zetten dat materiaal op authenticiteit kan controleren en ten tweede om te werken met een plicht voor alle procespartijen om slechts materiaal met een 'watermark' aan te leveren.

- Wat betreft de eerste optie bestaat nu het Nederlands Forensisch Instituut (NFI) die een rol speelt bij het analyseren van biologische, chemische en fysische sporen, expertise heeft omtrent digitale en biometrische sporen en van belang is bij bewijs in het kader van pathologie en toxicologie. Zo analyseert het NFI DNA-materiaal in strafzaken. Het NFI heeft al expertise op het gebied van digitaal forensisch onderzoek en het veiligstellen van digitale sporen. Overwogen kan dan ook worden om het NFI meer bevoegdheden en middelen te geven om ook digitaal bewijsmateriaal als zodanig op manipulatie en fabricatie te toetsen. Als alternatief zou een apart instituut kunnen worden opgericht dat zich hierop toelegt. Beide wegen zouden een flinke budgettaire post betekenen, zeker als inderdaad al het bewijs bij voorbaat op authenticiteit moet worden getoetst alvorens het in een rechtszaak mag worden ingebracht. Ook moet daarbij

worden bedacht dat zo'n instituut slechts 'waarheids-' of 'betrouwbaarheidspercentages' zal geven – bijvoorbeeld: de kans dat dit bewijs is gemanipuleerd is 29%. Tot slot is van belang dat het NFI sinds mei 2021 in samenwerking met de Universiteit van Amsterdam onderzoek verricht naar het herkennen van deepfakes.[411] De uitkomsten van dit onderzoek zouden kunnen worden afgewacht alvorens handen en voeten te geven aan deze reguleringsoptie.

- Daarnaast is de suggestie geopperd om slechts bewijs toe te laten als die van authenticiteitsbewijs wordt vergezeld. Deze reguleringsoptie vergt verdere uitwerking en zal slechts op de middellange termijn kunnen worden geïntroduceerd. Een voorbeeld dat werd geopperd is dat e-mails nu reeds zeer eenvoudig te manipuleren of fabriceren zijn en dat daar toenemende zorgen over zijn. Hierbij zou kunnen worden gewerkt met een mailsysteem dat een niet manipuleerbaar document van een of meerdere e-mails kan vervaardigen met een authenticiteitsstempel erop. Slechts dergelijke e-mails zouden dan als bewijsstuk in een rechtszaak kunnen worden toegevoegd. Dat zou vergen dat ofwel bestaande mailsystemen een dergelijke optie aanbieden ofwel dat er een plug-in of app op de markt verschijnt die dergelijks kan leveren. Naar analogie zou eenzelfde benadering kunnen worden gekozen bij ander digitaal materiaa. Dergelijke systemen zijn vooralsnog echter niet voor handen. Daarnaast is het aannemelijk dat als dergelijke systemen zullen worden ontwikkeld, ze zullen worden aangeboden door private partijen, wat tal van nieuwe vragen oproept over hun verantwoordelijkheid, de gekozen standaarden en hun betrouwbaarheid.

Reguleringsoptie 9:

- Bekijk in hoeverre er wet- en regelgeving of beleid dient te worden ontwikkeld of aangescherpt om (deep)fake bewijsmateriaal in de rechtszaal tegen te gaan

8.3 Handhaving en toezicht

Tot slot de handhaving van en het toezicht op de bestaande regels. Daarbij is duidelijk dat een keuze om slechts aanpassingen in het materieel recht te doen waarschijnlijk gelijk zal staan aan het accepteren dat de meeste schadelijke of mogelijk onrechtmatige deepfakes ongeadresseerd zullen blijven, om de redenen als genoemd in paragraaf 7.6. Hierin zijn onrechtmatige deepfakes niet uniek; dit geldt meer in algemene zin voor privacyschendingen in horizontale verhoudingen.

Vaak wordt, zowel in de politiek als daarbuiten, gesteld dat de oplossing zou zijn om de Autoriteit Persoonsgegevens meer middelen, mankracht en bevoegdheden te geven. Ook dat lijkt echter geen realistische oplossing. De samenleving sinds 1995 is fundamenteel veranderd; data-gedreven werken is de standaard voor vrijwel iedere organisatie. Bovendien hebben alle Nederlanders toegang tot digitale technieken; burgers verwerken vrijwel permanent persoonsgegevens van zichzelf en anderen. De Autoriteit kan niet alle content controleren op overeenstemming met de AVG, noch is zo'n superwaakhond wenselijk.

Wat betreft eventuele instrumenten om te zorgen voor een beter toezicht op en naleving van de regels liggen globaal drie instrumenten voor de hand: het aan banden leggen van de productie

en/of het aanbieden van deepfaketechnologie, het doen van ex-ante toetsten van concreet gebruik van deepfaketechnologie en het inzetten op bewustwordingscampagnes.

8.3.1 Verbieden

Het verbieden van technieken en producten ligt, terecht, gevoelig. Iedere techniek heeft immers positieve toepassingsmogelijkheden, het is juist in het veelvuldig gebruik van een techniek dat nieuwe, van tevoren onvoorziene mogelijkheden kunnen worden ontdekt en zowel burgers als bedrijven willen doorgaans dat technieken voor hen beschikbaar zijn. Dat zal ten aanzien van deepfakes niet anders liggen; burgers vinden het een leuke gadget om zichzelf en elkaar in allerlei grappige omstandigheden af te beelden, bedrijven vinden het leuk om hiermee te experimenteren, zien er diverse kansen in en mogelijkheden in termen van klantenbinding en customer-experience.

Toch is het verbieden van de technologie – in ieder geval voor om burgers toegang tot de techniek te ontzeggen – in dit geval een serieuze optie, omwille van de punten die in paragraaf 7.1 zijn genoemd. (1) De techniek wordt steeds beter, zodat deepfakes niet meer van echt te onderscheiden zullen zijn – daarin verschillen deepfakes ook wezenlijk van andere nep-content; (2) Technische mogelijkheden zullen vermoedelijk nooit meer dan 65% van de deepfakes er uit kunnen filteren en geven dan nog een 'waarheidspercentage'; (3) De techniek zal verder democratiseren– de verwachting is dat in het Westen binnen enkele jaren iedere burger toegang heeft tot de techniek; (4) Dit draagt bij aan het zogenoemde post-truth tijdperk – de voorspelling is dat over 6 jaar meer dan 90% van alle digitale content gemanipuleerd is; (5) Juist aan de democratisering van de techniek en de te verwachten grote hoeveelheden nepcontent die dat met zich zal brengen zitten grote maatschappelijke gevaren, zoals een teloorgang van de waarheid, problemen voor het functioneren van de journalistiek, de rechtspraak en de democratie en maatschappelijke onveiligheid voor vrouwen; (6) De meerwaarde van de techniek is gelegen in gebruik binnen specifieke professionele contexten, niet of nauwelijks in toepassingen door burgers; (7) Veruit de meeste use cases door burgers betreft het creëren van pornografisch materiaal over anderen zonder hun toestemming, zo'n 96%, waarbij ook nog moet worden gerekend op gebruik voor schadelijke misleiding, fraude en reputatieschade.

Het voorkomen dat burgers toegang hebben tot de techniek zou derhalve zeer veel maatschappelijke en persoonlijke schade voorkomen, terwijl het verlies slechts daarin gelegen zou zijn dat personen één instrument minder hebben om aan satire en vermaak te doen. De techniek zou dan – net zoals met sommige andere technieken en middelen is gebeurd – kunnen worden vrijgegeven voor bepaalde toepassingen of voor gebruik binnen bepaalde professionele contexten. Het verbod zou dan in ieder geval kunnen gelden voor deepfakes in horizontale verhoudingen, terwijl er ruimte kan worden gelaten voor de inzet van deepfakes door de politie, de filmbranch, de retailsector of andere professionele toepassingen al naar gelang dat wenselijk wordt geacht. Hierdoor zou slechts een veelvoorkomende positieve toepassing, namelijk de inzet van deepfakes voor satire, worden belemmerd en juist de grote maatschappelijke gevaren ten aanzien van het functioneren van de rechtstaat, de democratie en de media en de

maatschappelijke positie van (jonge) vrouwen worden verkomen. Die gevaren materialiseren zich immers niet of in veel mindere mate bij de inzet van deepfakes voor specifieke doeleinden binnen gecontroleerde settings.

Natuurlijk, zoals met ieder verbod, zal dit geen absoluut antwoord bieden op het gat tussen wet en praktijk, aangezien er burgers zullen zijn die de regels zullen omzeilen. Ook vuurwapens kunnen in Nederland op de zwarte markt worden verkregen en via duistere platforms op het internet worden besteld. Toch helpt zo'n verbod wel om te voorkomen dat dergelijke middelen op grote schaal in handen van burgers komen. Wel moet worden bedacht dat als het lukt om de verspreiding van deepfakes tegen te gaan, de manipulatie van online content evenwel door zal zetten. Toch wordt ook hier in ieder geval de meest vergaande vorm van manipulatie, met de meest verstrekkende gevolgen, tegengegaan.

Een dergelijk verbod zou op verschillende wijzen kunnen worden vormgegeven. De vraag daarbij is wie de primaire normadressant is.

Een verbod op de ontwikkeling van deepfaketechnologie zou een mogelijkheid kunnen zijn, maar de technologie of het ontwikkelen van deepfake-apps voor de consumentenmarkt gebeurt momenteel nauwelijks in Nederland zelf. Ook op Europees niveau zal dit weinig zoden aan de dijk zetten, omdat momenteel de meeste technologie in de Verenigde Staten wordt ontwikkeld. Een dergelijke verbod zou derhalve geen of weinig soelaas bieden, ook omdat het onwaarschijnlijk is dat in het buitenland gevestigde partijen zich iets gelegen zouden laten liggen aan dit verbod. Een meer voor de hand liggende optie zou zijn om de aanbieders van deepfaketechnologieën op de consumentenmarkt aan te spreken. Ook deze optie roept echter tal van vragen op. Gaat het dan om de grote App stores en zo ja, gebeurt dat middels co-regulering of een wetgevend kader, of gaat het om alle sites en netwerken die toegang tot de technologie geven en zo ja, hoe kunnen die worden genormeerd? Gaat het dan om apps of diensten die slechts en alleen een deepfake functionaliteit kennen of ook Apps en diensten waarbij deepfake een van de vele functionaliteiten is? En hoe zit het met open access software of deepfake-diensten op websites die in het buitenland worden gehost en gericht zijn op andere markten, maar die ook vanuit Nederland kunnen worden bereikt?

Eventueel zou er een regel kunnen komen voor accessproviders om burgers in Nederland geen toegang te geven tot sites, diensten en andere internetpagina's die deepfaketechnologie aanbieden of de productie van deepfakes mogelijk maken. De keuze om de regeldruk (deels) bij access providers te leggen wordt echter zelden gebruikt – kinderporno, is een voorbeeld, de toegang ontzeggen tot Piratebay is een ander voorbeeld. Deze voorbeelden laten direct ook zien hoe ongelijksoortig deepfaketechnologie is. In ernst en aard is het gebruik van deepfaketechnologie niet te vergelijken met kinderporno en het betreft niet één of enkele specifieke sites, zoals in het geval van Piratebay, maar om tientallen en op termijn waarschijnlijk honderden aanbieders. Aan de andere kant moet ook worden gesteld dat aangezien 96% van de deepfake toepassingen het genereren van non-consensual pornografisch materiaal betreft en dit zeer grote schade kan veroorzaken, een dergelijk route toch overwogen zou kunnen worden.

Tot slot zou de regeldruk bij de burger kunnen worden gelegd, bijvoorbeeld door een verbod op het downloaden, het in bezit hebben of het gebruik van deepfaketechnologie of technologie die voor het vervaardigen van deepfakes gebruikt kan worden. Hierbij moeten wel twee zaken worden opgemerkt. Ten eerste is de vraag hoe wenselijk het is om de reguleringsdruk bij de burger te leggen. Als een deepfake-app wel in een app store staat, dan gaat de gemiddelde burger er van uit dat het legitiem is om de app te downloaden en te gebruiken. Ten tweede is ook de vraag of een dergelijke verbod wel effectief zal zijn; van alle verboden op de toegang tot de techniek zal deze vermoedelijk het minst effect sorteren.

Ook moet worden bedacht dat een verbod op het produceren, aanbieden of gebruiken van deepfaketechnologie een goede en afgebakende definitie vergt. Daarbij moet een keuze worden gemaakt: wordt er een definitie gegeven van de technologie zelf en de functionaliteiten of van de producten die daarmee gemaakt kunnen worden. Als de eerste route wordt bewandeld, dan is de vraag of er op specifieke technologieën wordt ingezet (bijvoorbeeld slechts en uitsluitend GAN-technologie), waarmee veel potentiële andere technologie ongereguleerd wordt gelaten, of op meerdere technologieën (waarbij in ogenschouw moet worden genomen dat met zeer veel technieken fake-content kan worden geproduceerd, ook met bijvoorbeeld fotoshop). Als voor de tweede route wordt gekozen, dan is de vraag of alleen op videomateriaal wordt ingezet, of ook op audio, stilstaand beeld, tekst en andere signalen, zoals satellietsignalen. Gaat het dan om werkelijk iedere manipulatie, gaat het slechts om de meer significante manipulaties, en wat zou dan de grens moeten zijn, of gaat het om de intentie waarmee een deepfake wordt vervaardigd, en hoe definieer je die intentie dan? En hoe vertaal je een definitie van een deepfake dan terug naar een verbod op het genereren, aanbieden of gebruiken van een bepaalde technologie?

Reguleringsoptie 10:

Bekijk in hoeverre het wenselijk is om wet- of regelgeving te ontwikkelen waarin het vervaardigen, aanbieden, dowloaden of gebruiken van deepfaketechnologie of technologie die kan worden ingezet om deepfakes te genereren wordt verboden

8.3.2 Ex ante legitimiteitstoets

Ook zou er kunnen worden ingezet op een *ex ante* toets op de legitimiteit van specifieke deepfakes of gebruikstoepassingen van deepfakes. Hierdoor zou er niet (slechts) worden gekozen om de technologie of de toegang daartoe als zodanig te verbieden (zoals in de bovenstaande reguleringsoptie is besproken), maar het gebruik daarvan. Anders dan nu het geval is zou niet de controle op de legitimiteit van een deepfake achteraf, na publicatie op het internet of verspreiding onder grote groepen mensen, geschieden, maar vooraf. Een dergelijke plicht zou wederom kunnen worden belegd bij een aantal partijen.

Allereerst bij internetproviders die een mogelijke deepfake hosten of verspreiden. Steeds meer platforms verbieden in hun Terms and Conditions al bepaalde vormen van deepfakes. Een optie zou kunnen zijn om in overleg met de industrie dit verbod breed neer te leggen en ook afspraken te maken over de consequenties als gebruikers die regel overtreden, bijvoorbeeld schorsing van

het platform of dienst voor drie maanden bij de publicatie van een niet-schadelijke deepfake en schorsing voor een jaar bij de publicatie van een schadelijke deepfake. Ook kan er echter voor worden gekozen om de ontwikkelingen van dergelijke standaarden aan de industrie zelf over te laten, omdat zij zich op dit punt al lijken te bewegen. Wel zouden er andere vormen van (co-) regulering kunnen worden overwogen.

Zo kan er een specifieke plicht worden opgelegd aan providers om deepfake-detectietechnieken te gebruiken. Deze techniek filteren dan niet alle deepfakes eruit, maar wel een flink gedeelte. Daarbij zijn wel de volgende twee vragen van belang. Ten eerste: moet een bedrijf, als een detectiesysteem ziet dat een bepaalde video fake zou kunnen zijn, dergelijke content automatisch blokkeren of zijn sommige deepfakes wel toegestaan? Als voor de laatste optie wordt gekozen, dient de wetgever dan nadere duiding te geven aan welk type deepfakes wel of niet zijn toegestaan of wordt dat aan providers overgelaten, met als mogelijk gevolg dat verschillende providers verschillende regels hanteren? Ten tweede: mag hiermee worden gewerkt met een automatisch systeem of moet er menselijke tussenkomst zijn? Het vereiste van menselijke tussenkomst kan een zeer grote kostenpost voor dergelijke bedrijven opleveren en kan daarmee als onwenselijk worden beschouwd, maar is niet uniek, aangezien een bedrijf als Google ook veel heeft moeten investeren om alle 'vergeetverzoeken' te behandelen.

Het automatisch selecteren van content heeft drie belangrijke nadelen. Enerzijds dat zo'n systeem altijd zowel onder- als overinclusief is en anderzijds dat veel detectie-technieken een 'waarheidspercentage' geven en dit de vraag oproept bij welk 'waarheids-' of 'betrouwbaarheidspecentage' content moet worden toegelaten. Ook bij deze vraag speelt weer de keuze tussen het zetten van standaarden van overheidswege of dit aan de internetproviders te laten. Tot slot is het nadeel van zo'n automatisch systeem dat het mogelijk ook content blokkeert dat echt is, maar zo gemanipuleerd is dat er desalniettemin artefacten van deepfaketechnologie op achtergelaten zijn. Ook is op dit moment niet duidelijk in hoeverre de uiteindelijke Digital Service Act, die nu in ontwikkeling is, ruimte zal laten voor dergelijke regels op nationaal niveau.

Dan kan er ook nog worden gekozen voor additionele *ex ante* toetsen. Een dergelijke plicht kan bijvoorbeeld bij de Autoriteit Persoonsgegevens worden belegd. Hiertoe zouden burgers dan hun deepfake of een bepaalde gebruikstoepassing aan de AP moeten voorleggen, alvorens die te verspreiden of te publiceren; de AP zou dan de deepfake of toepassingsgebruik op overeenstemming met de AVG kunnen controleren. Hoe dit praktisch zou moeten plaatsgrijpen is echter lastig voor te stellen, aangezien het onwaarschijnlijk is dat alle burgers dit zullen doen en de AP zeker niet de mankracht (noch de wil) heeft om alle deepfakes op AVG-compliance te toetsen. Hoogstens kan een plicht voor de burger om deepfakes of gebruikstoepassingen aan de AP voor te leggen als praktische drempel gelden en een afschrikwekkende werking hebben. Dit zou dan met zich kunnen brengen dat burgers die weten of denken te weten dat hun deepfake niet in overeenstemming is met de AVG en dus worden afgewezen door de AP niet zullen vervaardigen. Toch is het sterk de vraag of burgers zich hier iets aan gelegen zullen laten liggen. Het zal de meesten volstrekt duidelijk zijn dat het genereren en verspreiden van pornografisch materiaal over een persoon zonder diens toestemming verboden

is. Of er dan een additioneel afschrikwekkend effect van een *ex ante* toets uitgaat valt te betwijfelen. Bovendien zal de AP de mankracht moeten hebben om na te gaan of gepubliceerde deepfakes eerder zijn voorgelegd aan de AP. Zodra de AP hier niet toe komt en burgers hiervan op de hoogte zijn, zullen ze in het vervolg waarschijnlijk ook geen deepfakes meer voorleggen aan de AP.

Datzelfde geldt voor een eventuele plicht voor de burger zelf om een dergelijk assessment uit te voeren. Wel is een dergelijke plicht redelijk eenvoudig op te leggen. De Autoriteit Persoonsgegevens kan bepalen dat verantwoordelijken voor de gegevensverwerking bij deepfakes standaard een Data Protection Impact Assessment (gegevenseffectbeoordeling) moeten doen en bij een geconstateerd hoog risico de AP daarvan op de hoogte moeten stellen en om advies over het al dan niet verspreiden van deepfakes moeten vragen. Dit zou het voordeel hebben dat een rechter of de AP niet iedere deepfake als zodanig op rechtmatigheid zal hoeven te beoordelen, maar zou kunnen volstaan met de constatering dat er geen DPIA is uitgevoerd en/of de AP niet om advies is gevraagd, wat een zelfstandige rechtsschending met zich brengt. Het is een procedurele/bureaucratische oplossing, die desalniettemin het handhaven van de materiele rechtsregels kan vergemakkelijken.

Reguleringsoptie 11:

Bekijk in hoeverre het wenselijk is om wet- of regelgeving aan te nemen die ziet op de verplichte controle door internetproviders, burgers en/of de AP van content op de authenticiteit en de rechtmatigheid voordat die content wordt gepubliceerd en/of verspreid

8.3.3 Bewustwording

Tot slot zou er kunnen worden ingezet op een bewustwordingscampagne. Dit kan op vier wijzen geschieden.

Allereerst is het van belang om via publieke campagne nieuwe normen en bijbehorende verboden voor het gebruik van deepfaketechnologie te markeren. Het belangrijkste hierbij is fake-porno, zeker met betrekking tot jonge vrouwen. Het kan met name jonge mannen, maar wellicht mannen in het algemeen, duidelijk worden gemaakt dat het vervaardigen en verspreiden van dergelijke content ontoelaatbaar en onrechtmatig is. Wellicht kan dit worden gecombineerd met een actie door het Openbaar Ministerie om een aantal extreme gevallen voor de rechter te brengen en daar breed media-aandacht voor te vragen.

Ten tweede zou er kunnen worden ingezet op het genereren van aandacht voor de positieve toepassingen van deepfaketechnologie. Zoals een voor deze studie geïnterviewde expert aangaf zet de voortdurende aandacht voor negatieve gebruikstoepassingen van deepfakes mensen ook op gedachten, terwijl andersom, voorbeelden van maatschappelijk wenselijke toepassingen en best practices bedrijven en burgers ook op ideeën zou kunnen brengen. Of de overheid hier een actieve rol in moet spelen of dat dit aan de industrie zelf kan worden gelaten is de vraag.

Ten derde zou kunnen worden gewezen op de mogelijkheden voor (potentiële) slachtoffers om zich te weren tegen deepfakes. Daarbij spelen globaal twee toepassingen een rol. Allereerst de pornografische deepfakes; hierbij zouden vrouwen moeten worden gewezen op

de bestaande mogelijkheden om deepfakes van het internet te verwijderen en op hun rechten worden gewezen. Er zou een landelijk meldpunt opgezet kunnen worden, waar slachtoffers van (pornografische) deepfakes naartoe kunnen gaan voor informatie en hulp. Wel gaven experts die voor deze studie zijn geïnterviewd aan dat hier moet worden gewaakt voor 'victim-blaming' en voor het privatiseren van een algemeen probleem bij specifieke slachtoffers, aangezien juist gewone vrouwen vaak niet de middelen en mogelijkheden hebben om lange rechtszaken uit te zitten. Ten tweede zijn er deepfakes die worden ingezet om valse informatie te verspreiden of om aan fraude en misleiding te doen. Burgers zijn slecht op de hoogte van deze gevaren en wapenen zich hier dus maar beperkt tegen. Een bewustwordingscampagne zou hier soelaas kunnen bieden, bijvoorbeeld door het adagium van één bron is geen bron ook aan burgers voor te leggen als mogelijke standaard bij het zien van sensationeel nieuws of als hen wordt gevraagd geld over te maken.

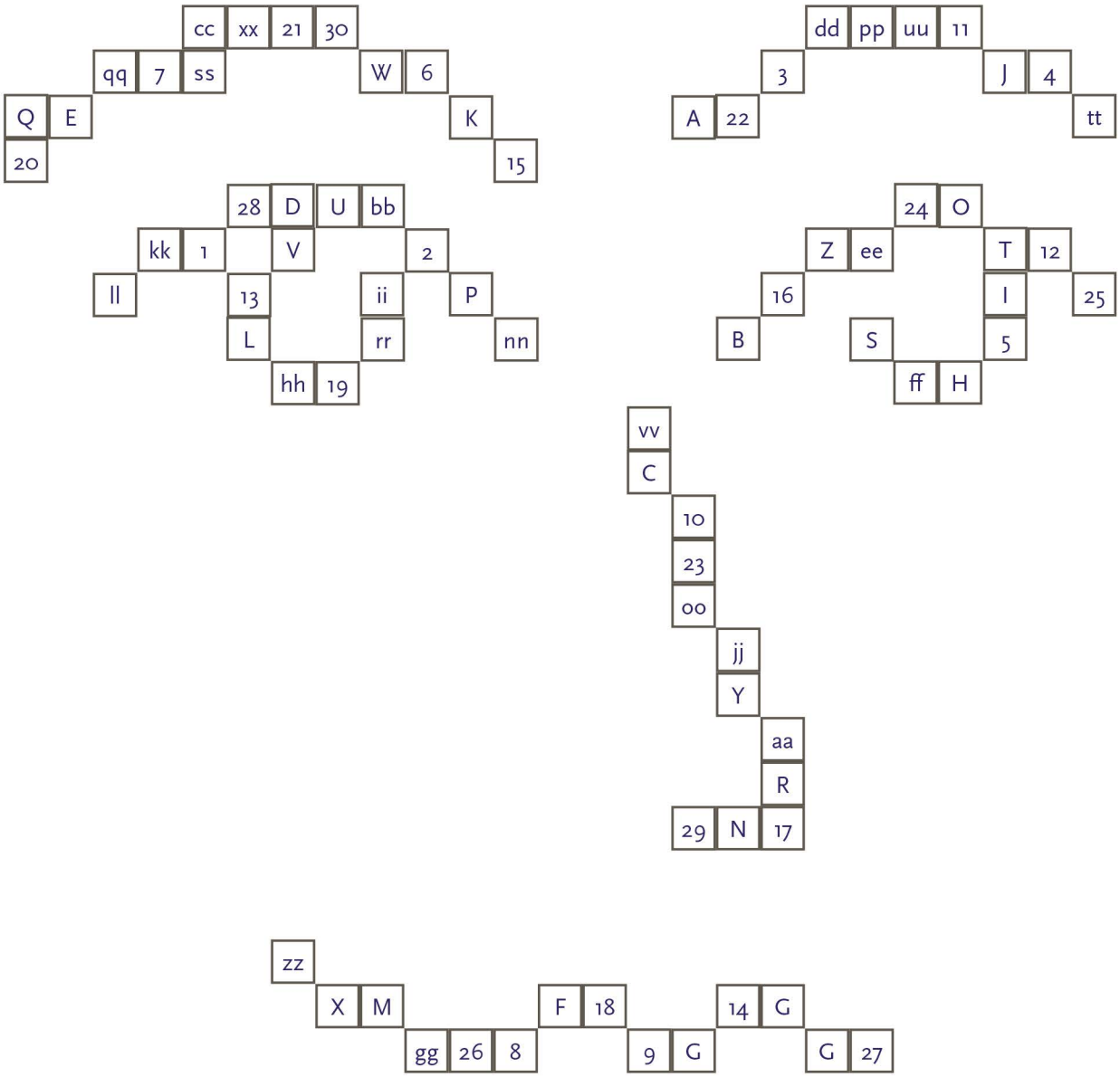
Ten vierde en tot slot zou een campagne zich kunnen richten op specifieke doelgroepen, waarbij journalisten en de zittende en staande macht het meest voor de hand liggen. Of de overheid hier een actieve rol in moet spelen of dat dit aan de beroepsverenigingen zelf kan worden overgelaten is de vraag. Wel is deze suggestie meermaals in de voor deze studie gehouden interviews gedaan. Daaruit en ook uit de bestudeerde literatuur blijkt dat er binnen de juridische praktijk nog weinig bekendheid is met het fenomeen deepfake, hier weinig tot geen procedures omtrent zijn ontwikkeld en de gedachte heerst dat het eenvoudig is om een deepfake te herkennen. Advocaten, het OM en rechters zouden zeer goed op de hoogte moeten zijn van de kans dat bewijs fake is en er zouden protocollen moeten worden ontwikkeld om bewijs op authenticiteit te controleren, net zoals nu reeds binnen de Nationale Politie geschiedt. Ook binnen de journalistiek zou een gezamenlijke strategie kunnen worden ontwikkeld.

Reguleringsoptie 12:

Start een publiekscampagne met informatie over de gevaren van deepfakes, waarin nieuwe sociale normen worden geëxpliciteerd en best practices worden benadrukt

Voetnoten hoofdstuk 8

- 404 Zie o.a.: Edwards, L., & Harbina, E. (2013). Protecting post-mortem privacy: Reconsidering the privacy interests of the deceased in a digital world. *Cardozo Arts & Ent. LJ, 32,* 83. Harbinja, E. (2017). Post-mortem privacy 2.0: theory, law, and technology. *International Review of Law, Computers & Technology, 31*(1), 26-42. Buitelaar, J. C. (2017). Post-mortem privacy and informational self-determination. *Ethics and Information Technology, 19*(2), 129-142. Lopez, A. B. (2016). Posthumous privacy, decedent intent, and post-mortem access to digital assets. Geo. Mason L. Rev., 24, 183. Gamba, F. (2020). The Right to be Forgotten and Paradoxical Visibility. Privacy, Post-privacy and Post-mortem Privacy in the Digital Era. *Problemi dell'informazione, 45*(2), 201-220. Holt, J., Nicholson, J., & Smeddinck, J. D. (2021, April). From Personal Data to Digital Legacy: Exploring Conflicts in the Sharing, Security and Privacy of Post-mortem Data. In *Proceedings of the Web Conference 2021* (pp. 2745-2756). Bikker, J. (2013). Disaster Victim Identification in the Information Age: The Use of Personal Data, Post-Mortem Privacy and the Rights of the Victim's Relatives. SCRIPTed, 10, 57. Davey, T. (2020). *Until Death Do Us Part: Post-mortem Privacy Rights for the Ante-mortem Person* (Doctoral dissertation, University of East Anglia).
- 405 <https://www.groene.nl/artikel/misogynie-als-politiek-wapen>.
- 406 <https://www.parool.nl/nieuws/man-37-uit-kudelstaart-opgepakt-voor-bedreiging-simons~be4608af/>.
- 407 <https://nos.nl/collectie/13860/artikel/2371839-vrouwelijke-oud-politici-het-zit-hem-vaak-in-taalgebruik>.
- 408 Zie echter: ECLI:NL:GHARL:2016:10242 Gerechtshof Arnhem-Leeuwarden, 21-000764-16, 20-12-2016.
- 409 Andere strategieën, zoals die aangaande Weerbaar Bestuur en Democratie, sluit hierop aan, en zie ook het voorstel over doxing. <https://www.internetconsultatie.nl/strafbaarstellinggebruikpersoonsgegevensvoorintimiderendedoeleinden/document/7177>.
- 410 Tweede Kamer, vergaderjaar 2019–2020, 30 821, nr. 91 en de Bijlage bij Kamerbrief beleidsinzet bescherming democratie tegen desinformatie.
- 411 <https://www.forensischinstituut.nl/actueel/nieuws/2021/05/22/uva-en-nfi-doen-onderzoek-naar-herkennen-deepfakes-en-verborgen-berichten-van-criminelen>.



Een deepfake wordt geconstrueerd op basis van al bestaande content, zoals foto's, video's en audio.
Daarom is het van groot belang om te weten waar de data vandaan komt, met welke intentie het is gemaakt of verzameld,
en wat de context is om te kunnen bepalen of de informatie betrouwbaar is.

9. Conclusie

Het onderzoek voor dit rapport betrof een verkennend onderzoek naar juridische aspecten van deepfakes in relatie tot horizontale privacy, dat wil zeggen de bescherming van de privacy in relaties tussen burgers onderling. De meest typische deepfake is een video die is gecreëerd met hoogwaardige technologische middelen waarin een bestaand persoon iets lijkt te doen of te zeggen dat diegene in werkelijkheid niet heeft gedaan of gezegd, waarbij het voor de consument van de video niet of nauwelijks mogelijk is deze manipulatie te ontwaren. Via literatuuronderzoek en juridische analyse, aangevuld met interviews en een landeninventarisatie, is een brede inventarisatie gemaakt van mogelijke lacunes of onvolkomenheden in de wetgeving en van reguleringsopties om deze lacunes of onvolkomenheden te adresseren.

Dit hoofdstuk geeft een samenvatting en de belangrijkste conclusies van het rapport. In paragraaf 9.1 worden kort de onderzoeksvragen herhaald. In paragraaf 9.2 volgen dan de belangrijkste inzichten uit het onderzoek, in paragraaf 9.3 het kader waarin de beantwoording van de onderzoeksvragen moet worden geplaatst en in paragraaf 9.4 worden de diverse reguleringsopties naast elkaar gezet en gekoppeld aan de onderzoeksvragen.

Dit hoofdstuk geeft een samenvatting en de belangrijkste conclusies van het rapport. In paragraaf 9.1 worden kort de onderzoeksvragen herhaald. In paragraaf 9.2 volgen dan de belangrijkste inzichten uit het onderzoek, in paragraaf 9.3 het kader waarin de beantwoording van de onderzoeksvragen moet worden geplaatst en in paragraaf 9.4 worden de diverse reguleringsopties naast elkaar gezet en gekoppeld aan de onderzoeksvragen.

9.1 Onderzoeksvragen

De probleemstelling waarmee dit onderzoek begon luidde: *'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?'*

Aan deze hoofdvraag is een aantal subvragen gekoppeld:

- 1. Welke uitingsvormen van deepfake zijn er te onderscheiden op basis van de bovenstaande indeling?
- 2. Hoe valt het maken en verspreiden van deepfakes binnen de huidige strafrechtelijke bepalingen?
- 3. Voldoet het huidige strafrecht om de makers van strafbare deepfakes aan te kunnen pakken?
- 4. Hoe valt het maken van een deepfake video binnen de huidige kaders van het gegevensbeschermingsrecht?
- 5. Welke mogelijkheden hebben burgers op dit moment om deepfakes van het internet te verwijderen en biedt dit hen voldoende handvatten?
- 6. Welke sanctie- of schadevergoeding mogelijkheden zijn er voor het onrechtmatig gebruik van persoonsgegevens voor het maken van deepfake video's?
- 7. Ligt het in de rede dat de Autoriteit Persoonsgegevens hier (ook) als handhaver op gaat treden, of ligt een strafrechtelijke benadering meer voor de hand?

- 8. In hoeverre biedt een vordering uit onrechtmatige daad mogelijkheid om content (die niet onrechtmatig is wegens inbreuk op gegevensbeschermingsecht) van het internet te verwijderen?
- 9. Hoe is het tegengaan van schadelijke of onrechtmatige deepfakes door andere landen gereguleerd voor wat betreft de diverse relevante rechtsgebieden en wat zijn daar eventueel reeds bekende voor- en nadelen van?
- 10. Dienen huidige en toekomstige onrechtmatige of strafbare uitingsvormen van deepfake technologie te leiden tot aanpassingen van de bestaande wetten en regels (met name AVG en strafrecht), of is bestaande wetgeving toereikend?
- 11. Indien aanpassingen worden voorgesteld: in hoeverre zorgen deze aanpassingen voor belemmeringen in de verdere ontwikkeling van bonafide deepfake toepassingen?

9.2 Inzichten

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. Negatieve toepassingen betreffen onder meer het genereren van (kinder)porno, fraude en misleiding, haat zaaien en aanzetten tot geweld, het verspreiden van misinformatie en het beïnvloeden van verkiezingen. Naast concrete gevolgen voor individuen kunnen deepfakes ook belangrijke maatschappelijke gevolgen hebben. Daarbij valt te denken aan een afnemend vertrouwen in de media, de politiek en de rechtsspraak en belemmeringen voor het functioneren van deze instituten door de hoeveelheid nepmateriaal dat wordt gecreëerd en verspreid. Ook hebben de deepfake pornotoepassingen niet zelden een negatief effect op vrouwen en hun maatschappelijke positie.

Uit de juridische analyse van het vigerende materieelrecht en de toepasbaarheid daarvan op deepfakes in horizontale verhoudingen blijkt dat het Nederlandse strafrecht over het algemeen goed is 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'.

Verder blijkt dat deepfakes tegen een aantal obstakels oplopen onder het gegevensverwerkingsregime van de Algemene Verordening Gegevensbescherming. Er moet een legitieme verwerkingsgrond zijn. Allereerst kan de verwerker kiezen 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. Dat geldt ook 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. Wel moet worden bedacht dat er uitzonderingen kunnen bestaan in de vorm van de huishoudelijke exceptie en de verwerking van gegevens in het kader van de vrijheid van meningsuiting. Hoe nauw of wijd deze uitzonderingen dienen te worden geïnterpreteerd in de context van deepfakes is echter niet op voorhand en in het algemeen vast te stellen. Belangrijk is dat de AI-regulering, zoals voorgesteld door de Commissie, wel een specifieke bepaling bevat aangaande deepfakes, maar dat die slechts een transparantieverplichting betreft die voortbouwt op wat reeds volgt uit de AVG.

Voorts moet binnen het Europees Verdrag voor de Rechten van de Mens 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, goede naam en reputatie. Ook heeft het Hof bepaald 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, beoordeeld, kan niet in algemene zin worden gezegd hoe deze

twee rechten zich bij deepfaketoepassingen tot elkaar verhouden. Dit zal per zaak moeten worden bekeken. Wel zijn twee bijzondere punten van belang. Enerzijds heeft het Hof een uitgebreide doctrine aangaande wat het noemt de 'reasonable expectation of privacy'; daaruit volgt dat mensen ook in een werkomgeving of in de publieke ruimte mogen verwachten dat hun privacy wordt beschermd. Zelfs als een persoon extreme seksuele afbeeldingen van zichzelf op het internet zet, dan nog mag hij verwachten dat zijn privacy door anderen wordt gerespecteerd. Dat is van belang omdat hieruit volgt dat het recht op privacy op veruit het meeste materiaal dat wordt gebruik voor het genereren van deepfakes van toepassing zal zijn. Anderzijds geldt er een speciale doctrine voor bekende personen. Dit is van belang omdat de meeste deepfakes worden gemaakt ofwel over personen in de directe omgeving van de maker ofwel over bekende personen. Deze bekende personen moeten, zeker als het politici of ambtsdragers betreft, volgens het Hof meer dulden in termen van inperkingen in hun privésfeer en hun reputatie, eer en goede naam dan gewone burgers. Toch heeft het Hof eveneens benadrukt dat dergelijke inperkingen alsnog proportioneel dienen te zijn en dat ook publieke personen een recht op privacy toekomt.

Ten aanzien van het procesrecht blijkt dat er zowel binnen het civiel recht als binnen het strafrecht een complex stelsel aan regels, indicaties en contra-indicaties speelt bij mogelijke bewijsvraagstukken in het kader van deepfake video's, afbeeldingen, audiofragmenten of ander materiaal. Zowel het burgerlijk procesrecht als het strafprocesrecht zijn redelijk open van aard. Er bestaan geen bijzondere bepalingen ten aanzien van deepfakes. Net zoals bij de materieelrechtelijke bepalingen, zijn de procesrechtelijke regels op zich breed genoeg om vraagstukken omtrent de authenticiteit van deepfakes te adresseren. In die zin zijn deepfakes slechts de zoveelste variant van technische mogelijkheden om bewijsmateriaal te manipuleren of te fabriceren. Toch geldt ook hier dat zowel de ogenschijnlijke echtheid en de toegang tot dergelijke middelen door iedere burger een wezenlijk risico vormen voor juridische processen.

Uit dit rapport blijkt verder dat een van de belangrijkste knelpunten in de horizontale privacybescherming in het algemeen de handhaving van en het toezicht op de materiële rechtsregels is. Omdat veel van de regels momenteel zien op het gebruik van technologieën en niet op het produceren of aanbieden van technologieën, omdat doorgaans pas wordt geverifieerd of materiaal aan de wettelijke regels voldoet nadat dat is vervaardigd en verspreid en omdat het nagaan of bepaald materiaal al dan niet rechtmatig is, enkele uitzonderingen daargelaten, bij de burger is belegd, is er een praktijk ontstaan waarin de geldende regels veelvuldig worden overtreden, zonder dat hierop wordt geacteerd.

Uit de landenstudie blijkt een diversiteit aan benaderingen van deepfakes en daaraan gerelateerde onderwerkpen als des- en misinformatie. Zo kiest Australië voor een zelfreguleringsinstrument van de grote internetondermemingen om misinformatie aan te pakken, wordt er binnen de Canadese context met name nadruk gelegd op het gebruik van deepfakes binnen de politieke context, wordt in Duitsland met name gekeken naar de mogelijkheid om deepfakes in te zetten voor politieinfiltratie in criminele netwerken, wordt in de Filippijnen de nadruk gelegd op de bescherming van de

individuele privacy en identiteit van burgers, richt Frankrijk zich op nepnieuws, onderstreept men in India het belang van een veilige en betrouwbare digitale infrastructuur, is er in Oekraïne een wetsvoorstel aanhangig die tot doel heeft de verspreiding van desinformatie tegen te gaan, onderkent Singapore het mogelijk schadelijke effect van manipulatie en desinformatie op onder meer de nationale veiligheid, sociale contacten en democratische verkiezingen en zal binnen het Verenigd Koninkrijk de Online Harms Bill op termijn nieuwe aanknopingspunten bieden om mensen te helpen veilig toegang te krijgen tot het internet. Twee landen hebben specifieke regelgeving aangenomen omtrent deepfakes, namelijk China en de Verenigde Staten.

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, onder meer door: a) de installatie van controle- en identificatietechnologieën, b) het stopzetten van de doorgifte van illegale of onwettige informatie, het nemen van verdere maatregelen zoals het wissen en voorkomen van verspreiding, het bijhouden van desbetreffende gegevensbestanden, enz. en c) het melden van dergelijke inhoud aan regelgevende instanties op het gebied van netwerken, cultuur en omroep. 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 – Californië, Virginia en New York – tot dusver wetgeving vastgesteld om het probleem van het zonder toestemming creëren en verspreiden van expliciet seksueel materiaal aan te pakken. 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 wettelijke privaatrechtelijke grond tot het instellen van een vordering biedt. Sectie 1708.86 van het Californische burgerlijk wetboek bepaalt dat een persoon die als gevolg van "digitalisering" een voorstelling lijkt te geven die hij niet daadwerkelijk heeft gegeven, of die een "gewijzigde afbeelding" geeft, een vordering kan instellen tegen (i) degenen die expliciet seksueel materiaal creëren en opzettelijk openbaar maken, en die weten of redelijkerwijs hadden moeten weten dat de persoon die in het materiaal wordt afgebeeld, niet heeft ingestemd met de creatie of openbaarmaking ervan; en (ii) eenieder die expliciet seksueel materiaal opzettelijk openbaar maakt, wetende dat het is gecreëerd zonder de toestemming van de persoon die erin wordt afgebeeld. 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. Het ongeoorloofd gebruik van een "digitale kopie" is een grond voor het instellen van een rechtsvordering. De wet voorziet in rechtsmiddelen in geval van ongeoorloofd gebruik van de beeltenis van een overleden uitvoerende kunstenaar in een artistiek werk dat met behulp van deepfaketechnologie is vervaardigd. Verder hebben enkele staten 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 elf interviews gehouden met internationale experts en vier interviews met Nederlandse experts op het gebied van het procesrecht. De algemene verwachting van deze experts is dat het gebruik van deepfaketechnologie de komende jaren een grote vlucht zal nemen. Daarbij speelt dat steeds meer technieken, of misschien beter gezegd, informatietechnologie per definitie, de waarheid manipuleert. Dat gaat vaak om relatief kleine manipulaties: fotocamera's die rode ogen genereren, video-beldiensten die de huid van een persoon egaler laten lijken, audio die middels compressie een deel van zijn kwaliteit en een aantal van de hogere geluidsregisters verliest. Experts 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 een bestaand en niet gemanipuleerd materiaal sporen (artefacten) van manipulaties achter te laten, die de detectie-technologie kan ontdekken. De detectie-technologie zal het materiaal dan aanmerken als fake en 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 is, dat wil zeggen niet gemanipuleerd, is 78%. Het is dan lastig te bepalen hoe je deze video zou moeten behandelen.

De experts wijzen er bovendien op dat de wetenschap dat een filmpje deepfake is maar van beperkt belang is. De sociale consequenties van een porno-filmpje voor een opgroeiend meisje kunnen aanzienlijk zijn, ook al weet de groep dat het een deepfake betreft. Ook kan het zien van een dergelijk filmpje het zelfbeeld van de vrouw in kwestie aantasten. Ook al weet ze dat het materiaal nep is, toch kan het bekijken van jezelf terwijl je allerhande expliciete handelingen verricht een negatieve impact hebben op je zelfvertrouwen en zelfwaarde. Dit punt – dat ook al is het bekend dat bepaalde content fake is, de gevolgen er niet minder om zijn – geldt bijvoorbeeld ook bij fakenieuws. Berichten die op het eerste gezicht zeer dubieus lijken worden maar al te graag omarmd en gedeeld binnen groepen waarbinnen het bericht aansluit op hun wereldbeeld en/of politieke overtuiging. Ook als later het bericht wordt ontkracht, dan nog houdt zo'n groep vaak vol dat het specifieke bericht dan wel nep was, maar dat de onderliggende waarheid die daarvan uitging wel degelijk zou kloppen. Bij nepnieuws dat onder het grotere publiek wordt verspreid, maar later wordt ontkracht, is de vrees dat het aanvankelijke (vaak sensationele) fake-bericht significant meer aandacht zal genereren dan de latere nuance of ontkrachting van dat bericht. Bovendien blijft er dan vaak alsnog een 'er was toch iets met...?'-gevoel achter.

9.3 Kader

Hoe moeten we in het licht van bovenstaande inzichten nu aankijken tegen deepfakes? 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, andere 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.

9.4 Reguleringsopties

Tot slot heeft deze studie een breed palet opgeleverd van mogelijke reguleringsopties. Sommige daarvan zijn direct invoerbaar en concreet uitgewerkt, andere betreffen meer opties voor de lange termijn en vergen structurele wijzigingen in het Nederlandse rechtstelsel. De reguleringsopties moeten in onderlinge samenhang worden bezien. Meerdere opties kunnen een onderliggend probleem adresseren; als er wordt gekozen voor de ene optie, kunnen andere dan achterwege blijven. Ook zijn diverse combinaties van reguleringsopties denkbaar. We noemen de reguleringsopties als een breed palet aan mogelijkheden voor regulering. Op basis van dit verkennende onderzoek is het niet mogelijk om te stellen wat er precies moet gebeuren; daarvoor is nader onderzoek nodig naar de precieze vormgeving en te verwachten gevolgen van specifieke interventies, alsook een beleidsmatige afweging tussen voor- en nadelen daarvan. Het is aan de wetgever en beleidsmakers om keuzes te maken binnen het brede palet aan reguleringsopties.

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 aangescherpt om (deep)fake bewijsmateriaal in de rechtszaal tegen te gaan	Reguleringsoptie 10: Bekijk in hoeverre het wenselijk is om wet- of regelgeving te ontwikkelen waarin het vervaardigen, aanbieden, dowloaden of gebruiken van deepfaketechnologie 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 gepubliceerd 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 misleidende uitingen		Reguleringsoptie 12: Start een publiekscampagne 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 ontwikkelen ten aanzien van het beïnvloeden van verkiezingen of politieke besluitvorming middels het creëren of verspreiden van misinformatie		

Figuur 19: Overzicht van reguleringsopties

De belangrijkste uitkomst van dit onderzoek, die ook aansluit bij de drie eerdere voor het WODC verrichte studies naar horizontale privacy,[412] is dat de meeste problematische vormen van deepfakes reeds goed gereguleerd zijn. Het vervaardigen van pornografisch materiaal van een ander zonder diens toestemming is al verboden; het genereren van kinderporno van een fictief kind is al verboden; het plegen van fraude en misleiding middels een deepfake is al verboden; het aandragen van valselijk bewijsmateriaal in een rechtszaak is al verboden; het aanzetten tot haat of geweld tussen groepen middels een deepfake is al verboden; het zonder toestemming exploiteren van iemands beeld of gelijkenis of creatieve werken is al verboden; het schade berokkenen aan een ander middels een fake-bericht kan al onder het onrechtmatige-daadsregime worden aangepakt; etc. De juridische inkadering van deepfakes is daarom niet het primaire probleem, het probleem is de handhaving van de bestaande en eventuele aanvullende rechtsregels.

Uit deze studie en de voorgaande drie studies over horizontale privacy blijkt dat het investeren in verdere handhaving door de AP en/of het OM slechts (zeer) gedeeltelijk solaas kan bieden. Enerzijds verschijnt er zoveel materiaal (en zal op termijn naar verwachting zo'n 90% daarvan enige vorm van manipulatie hebben ondergaan) dat het onmogelijk is om alle gepubliceerde content op authenticiteit en/of legitimiteit te controleren. Hoe groot de capaciteit van de AP en/of het OM ook wordt gemaakt, deze instituten zullen nimmer genoeg capaciteit hebben om alle problemen te adresseren. Anderzijds is de vraag of een superwaakhond die van overheidswege alle publicaties van burgers nauwgezet volgt en beoordeelt wenselijk is. De kuur kan dan erger zijn dan de kwaal, zo werd ook al in de vorige studies over horizontale privacy geconcludeerd.

Iedere burger heeft weliswaar allerhande proces- en klachtrechten, zoals bijvoorbeeld onder de AVG en het onrechtmatige-daadsrecht, maar het is duidelijk dat gezien de hoeveelheid partijen die gegevens over een gemiddelde burger verwerken en de complexiteit van een rechtsgang, deze middelen weinig soelaas bieden in de data-gedreven samenleving. Dit probleem is uitgebreid onderzocht en besproken in een eerdere studie verricht voor het WODC.[413] Daaruit bleek onder meer dat het lang niet altijd duidelijk is voor iemand dat zijn data worden of zijn verzameld of dat er een deepfake van hem is verspreid op het internet (bijvoorbeeld op een pornosite of Geenstijl.nl). Zelfs als hij dat wel weet of te weten komt, dan nog is niet altijd duidelijk wie verantwoordelijk kan worden gehouden of aansprakelijk kan worden gesteld. Om achter de identiteit van de dader te komen is vaak de medewerking van internet intermediairs noodzakelijk, maar die willen niet altijd meewerken (zonder last van de rechter) vanwege de privacybelangen van degene die het materiaal heeft geplaatst. Dat betekent dat er vaak twee rechtszaken nodig zijn, één om de identiteit van de dader te achterhalen en een ander om de dader in rechte aan te spreken. Als het daarbij nog gaat om een verwijderverzoek ten aanzien van het platform of ten aanzien van eventuele kopieën die elders zijn gepubliceerd kan soms een derde, vierde en volgende rechtszaak nodig zijn. Dit vergt tijd, geld en energie die burgers vaak ontberen; de schadevergoeding die wordt geboden als de burger in het gelijk wordt gesteld is doorgaans verwaarloosbaar, waardoor veel mensen afzien van rechtszaken bij privacyschendingen, tenzij het om zeer ernstige

gevallen gaat. Tot slot speelt voor burgers ook dat een rechtsgang meer aandacht kan genereren voor het vermeend onrechtmatige materiaal.

Een eerste inzicht is derhalve dat het primaire probleem van deepfakes in het bijzonder en privacyschendingen in horizontale relaties in het algemeen niet in de materieelrechtelijke regulering is gelegen, maar in het toezicht op en de handhaving van de regels. Een tweede is dat zowel van overheidsinstanties als van burgers moeilijk kan worden verwacht dat zij alle (mogelijk illegitieme) content op authenticiteit en/of legitimiteit controleren. Daarbij speelt dat het huidige recht primair uitgaat van zogenoemde *ex post* regulering: de ontwikkeling en het aanbieden en de toegang tot de techniek wordt niet aan banden gelegd, maar het gebruik daarvan voor specifieke doeleinden; dat gebruik wordt doorgaans pas nadat het materiaal is gemaakt, verspreid en in de openbaarheid gekomen op rechtmatigheid getoetst. Hierdoor en door het veelvuldig gebruik door burgers van allerhande producten en diensten waarmee gegevens over anderen worden verwerkt is het aantal 'alledaagse' privacyschendingen de laatste jaren exponentieel toegenomen. Door de beperkingen in opsporings- en handhavingscapaciteiten heeft er een normalisering van kleine privacyschendingen plaatsgegrepen. Ten derde speelt een rol dat de meerwaarde van deepfakes in horizontale relaties momenteel beperkt is tot satire, terwijl meer dan 95% van de momenteel gepubliceerde deepfakes duidelijk strafrechtelijk verboden is (het gaat dan met name om het genereren van deepfake pornobeelden van anderen zonder hun toestemming); de meerwaarde van de inzet van de technologie is primair gelegen in professionele toepassingen. Uiteraard valt niet uit te sluiten dat dit in de toekomst anders zal zijn, maar desgevraagd konden de voor deze studie geïnterviewde experts weinig andere mogelijke positieve toepassingen door burgers in hun privéhoedanigheid naar voren brengen. Ten vierde is duidelijk dat de grote maatschappelijke problemen (problemen ten aanzien van onder meer de rechtstaat, betrouwbare nieuwsvoorziening en het democratische proces) met name samenhangen met de (te verwachten) grote hoeveelheid gemanipuleerde content gegenereerd door burgers en niet met de inzet van deepfakes in afgebakende situaties door professionele partijen.

Daarom kan in het kader van deepfakes worden overwogen om met *ex ante* regulering te werken, ofwel door het produceren, aanbieden, gebruiken of in bezit hebben van deepfake-technologie te verbieden voor de consumentenmarkt (reguleringsoptie 10) ofwel door een verplichte *ex ante* legitimiteitstoets in te voeren, die dient moet worden uitgevoerd voordat materiaal wordt gepubliceerd en/of verspreid door burgers (reguleringsoptie 11). Zulke *ex ante* regels zouden het gebruik van deepfaketechnologie door professionele partijen ongemoeid laten. Een *ex ante* toets kan ook worden verplicht voordat materiaal mag worden gebruikt voor strafrechtelijk onderzoek of mag worden ingebracht in de rechtszaal als wettig bewijs (reguleringsoptie 9). Dat dergelijke regels moeten worden ingevoerd om de handhavingsdruk ten aanzien van horizontale privacyschendingen in het algemeen en de inzet van deepfakes in horizontale relaties in het bijzonder te verminderen ligt meer voor de hand dan hoe dit dient te geschieden. Er zijn diverse potentiële normadressaten mogelijk voor dergelijke *ex ante* regels; elke keuze heeft zijn eigen merites en bezwaren. Hier zal nader onderzoek naar moeten

worden verricht en een politieke keuze in worden gemaakt. Er kan op dit punt ook worden gedacht aan een combinatie van regels, zoals een verbod om deepfake-technologie of -applicaties op de consumentenmarkt aan te bieden en een plicht aan websites en platforms om content te monitoren. Natuurlijk moet worden bedacht dat dergelijke regels kunnen worden omzeild, maar dat geldt voor iedere regel. Hoe dan ook helpt een dergelijk verbod om te voorkomen dat burgers op grote schaal deepfakes produceren, wat een gunstig effect zal hebben op het aantal deepfakes, wat daarmee de handhavingsdruk sterk zal doen verminderen. Tot slot kunnen ook bewustwordings- en publiekscampagnes helpen bij het informeren van burgers over juridische en maatschappelijke regels (reguleringsoptie 12). Toch is de onbekendheid met de juridische regels niet de primaire oorzaak van de veelvuldige wetsovertredingen (het mag bijvoorbeeld als gevoeglijk bekend worden beschouwd dat een fake pornofilm van iemand anders zonder diens toestemming maken en publiceren niet is toegestaan); daarom zal het informeren van burgers over vigerend recht het handhavingsprobleem als zodanig niet wegnemen.

Daarnaast is een aantal punten geïdentificeerd dat als juridische lacune zou kunnen worden bestempeld. De meeste duidelijke lacune is dat misbruik van biometrische gegevens in gevallen waarin die gegevens niet identificatie tot doel hebben momenteel niet strafbaar is, omdat zulks tussen wal (art. 231a Sr) en schip (artikel 231b Sr) valt. Dit valt te verhelpen door een tekstuele aanpassing aan artikel 231b Sr (reguleringsoptie 2). Bij andere punten dient er eerst een maatschappelijk en politiek debat plaats te hebben omtrent de vraag of er momenteel een juridische lacune is – een vraag die verschillend kan worden beantwoord al naar gelang de ethische, politieke of normatieve stroming die men aanhangt – en in hoeverre juridisch ingrijpen wenselijk is, onder meer omdat zulk ingrijpen nadelige effecten kan hebben, bijvoorbeeld voor de vrijheid van burgers.

Een voorbeeld is het feit dat het in de privésfeer verwerken van persoonsgegevens over anderen momenteel noch strafrechtelijk noch middels de AVG is gereguleerd. Dit kan problematisch zijn, bijvoorbeeld als een burger van een ander een deepfake pornofilm maakt en zichzelf daar thuis aan verlekkert. Daarom kan worden overwogen om zulks strafrechtelijk (reguleringsoptie 1) en/of middels het gegevensbeschermingsrecht (reguleringsoptie 5) te adresseren. Toch roept dit ook nieuwe reguleringsvragen op. Hoe moet de Autoriteit Persoonsgegevens of het Openbare Ministerie zicht houden op wat mensen in hun privésfeer doen? In hoeverre is het überhaupt wenselijk om privégedragingen juridisch in te kaderen? En als de AP en/of het OM in de privégedragingen van burgers gaan treden, is de kuur dan niet erger dan de kwaal? Het vergt een politiek en maatschappelijk debat alvorens nadere regels op dit punt worden aangenomen.

Datzelfde kan worden gezegd over de mogelijke regels die online uitingen aan banden leggen. Het is duidelijk dat deepfakes een nefaste invloed kunnen hebben op het functioneren van de media, dat fakenews democratische verkiezingen kunnen beinvloeden en dat deepfakes over publieke personen er aan kunnen bijdragen dat gekwalificeerde mensen afzien van een publieke functie. Daarom zou kunnen worden bekeken in hoeverre nadere regels kunnen worden gesteld aan uitingen over publieke personen, die nu

meer moeten dulden dan privépersonen, om hen zodoende meer bescherming te bieden tegen ongewenste en onware uitingen door burgers (reguleringsoptie 6) en in hoeverre de strijd kan worden aangebonden tegen de verspreiding van nepnieuws (reguleringsoptie 7) en de beïnvloeding van democratische verkiezingen of politieke besluitvorming door middel van nepnieuws (reguleringsoptie 8).

Een additioneel argument voor de introductie van nadere regels op dit punt is dat momenteel niet alle kwalijke uitingen onrechtmatig zullen zijn aangezien het doen van onware uitingen niet an sich is gereguleerd. Een onware, onjuiste of misleidende uiting kan onder het huidige regime wel worden geadresseerd, maar slechts als er schade is ontstaan, bijvoorbeeld aan persoonlijke belangen (onder het onrechtmatige-daadsregime) of aan bepaalde maatschappelijke belangen, zoals wanneer haat tegen bepaalde groepen wordt aangewakkerd (onder het strafrecht). Dit roept globaal drie problemen op. Ten eerste kan het moeilijk zijn om het causale verband aan te tonen tussen een onware, onjuiste of misleidende verklaring en de (voorzienbare) schade die daarmee wordt veroorzaakt (bijvoorbeeld de haat die een deepfake heeft opgewekt tegen minderheidsgroepen). Ten tweede kunnen onware, onjuiste of misleidende uitingen problematisch zijn omdat ze de grens tussen feit en fictie doen vervagen, zelfs als ze geen concrete schade aanrichten. Ten derde zijn er onware, onjuiste of misleidende uitingen die wel schade berokkenen, maar die zeer moeilijk te koppelen zijn aan specifieke wettelijke bepalingen. Er kunnen bijvoorbeeld nep-satellietbeelden worden geproduceerd waarin Rusland zijn kernraketten naar de Letse grens lijkt te verplaatsen, waardoor er politieke spanningen ontstaan. Of er kan nepnieuws worden verspreid over Covid-vaccins, wat leidt tot een afname van de vaccinatiebereidheid. Of een politieke leider kan een video verspreiden, waardoor het lijkt alsof er duizenden supporters bij zijn bijeenkomsten aanwezig zijn, terwijl er slechts een handvol staan. Nieuwe regels kunnen meer helderheid scheppen in de rechtmatigheid van dit soort uitingen.

Toch geldt ook hier dat nadere regels over uitingen aangaande publieke personen of onware uitingen al dan niet tijdens democratische verkiezingen onwenselijk kunnen worden geacht te zijn omdat ze fnuikend kunnen zijn voor het vrije en open debat, dat juist zo essentieel is voor een democratische rechtsstaat. Ook is het sterk de vraag in hoeverre het wenselijk is als de overheid zich gaat bemoeien met wat waar is en wat niet; juist dit was de vrees van Orwell. Alhoewel op ieder van deze aarzelingen een juridische oplossing kan worden gevonden, is ook hier van belang dat een maatschappelijk en politiek debat wordt gevoerd alvorens er nadere regels worden geïntroduceerd.

Tot slot heeft deze studie twee nieuwe(re) vraagstukken geïdentificeerd ten aanzien waarvan nadere regels wenselijk kunnen zijn, maar waarvan de beantwoording van wat zou moeten zijn toegestaan en wat niet wederom afhangt van politieke voorkeur en/of de ethische stroming die wordt aangehangen. Het betreft hier de deepfakes van overleden personen en deepfakes van volledig fictieve personen.

Overleden personen kunnen weer tot leven worden gebracht; momenteel staat daar juridisch weinig aan in de weg (behalve wellicht in sommige zaken het intellectueel eigendomsrecht). Dit

roept tal van vragen op zoals: Is het wenselijk en toegestaan om overleden historische figuren op scholen les te laten geven? Is het wenselijk en toegestaan om overleden kunstenaars een rondleiding te laten geven in een museum? Is het wenselijk en toegestaan om overleden acteurs in films te laten figureren? Is het wenselijk en toegestaan om een overleden persoon in een pornofilm te laten spelen? Is het wenselijk en toegestaan om overleden artiesten nog concerten te laten geven? Is het wenselijk en toegestaan om dagelijks te blijven communiceren met een deepfake van de overleden echtgenoot? Is het wenselijk en toegestaan om iemand via een realistische deepfake een speech te laten geven op zijn eigen begrafenis zonder dat hij daarvoor toestemming heeft gegeven? Om over deze en andere vragen meer duidelijkheid te geven zou aparte wetgeving over post-mortum privacy kunnen worden vervaardigd (reguleringsoptie 3).

Dat geldt ook voor de vele vragen die spelen ten aanzien van het maken en gebruiken van volledig fictieve, maar zeer realistische, door deepfaketechnologie gegenereerde personen. Vragen die hier spelen zijn onder meer: Mag de politie door middel van een fake persoon infiltreren in een crimineel netwerk, door middel van fakekinderporno pederasters opsporen of middels fake klantaccounts vrouwenhandelaren in kaart brengen? Mogen fictieve personen worden ingezet voor therapeutische doeleinden en zo ja, gelden daarvoor morele grenzen? Mogen fictieve personen worden ingezet voor het maken van (porno)films en zo ja, gelden daarvoor morele grenzen? Ook op dit punt zou nadere regelgeving kunnen worden overwogen (reguleringsoptie 4), maar geldt dat de beantwoording van de vragen niet uit een juridisch onderzoek als deze kan voortvloeien.

Voetnoten hoofdstuk 9

412 E. Keymolen, M. Noorman, B. van der Sloot, C. Cuijpers, B.J. Koops & B. Zhao, Op het eerste gezicht: Een verkenning van gezichtsherkenning en privacyrisico's in horizontale relaties, Tilburg: TILT 2020, Projectnummer: 2992. M. Galič, M. Noorman, B. van der Sloot, B.J. Koops, C. Cuijpers, R. Gellert, & T. van Delden, Spioneren met hobbydrones en andere technologieën door burgers: een verkenning van de privacyrisico's en reguleringsmogelijkheden, Tilburg: TILT WODC 2020, Projectnummer: 3063. B. W. Schermer & B. van der Sloot, Het recht op privacy in horizontale verhoudingen, Amsterdam: Considerati WODC 2020.

413 B. van der Sloot & S. van Schendel, 'De Modernisering van het Nederlands Procesrecht in het licht van Big Data: Procedurele waarborgen en een goede toegang tot het recht als randvoorwaarden voor een data-gedreven samenleving', WODC 2019.

10. Bijlagen

10.1 Landenstudies

10.1.1 China

Author: **Bo Zhao**[414]

1. Introduction

With a dynamic and still fast growing social media market, Chinese developers and users have created multiple innovative ways to make use of new audio/video technologies on the internet. When big players like Douyin, the Chinese version of TIkTok, and Taobao (part of Alibaba) rolled out deepfake technologies to attract more users to their services, China has become one of the world's deepfake capitals.[415] Like the rest of the world, China has been suffering from escalating negative impacts of the quick roll out of deepfake technologies e.g., audio-video based software, AI, and new network technologies and applications,[416] in particular regarding public safety, legitimate rights and interests of consumers, law enforcement, market order, etc,417 in addition to disinformation and fake news which are of political significance.[418] As a Chinese official spokesman made it clear, deepfakes:

> "may be used to engage in activities prohibited by laws and regulations that endanger national security, undermine social stability, disrupt social order, and infringe on the legitimate rights and interests of others, causing political security risks, national security and public security risks and adversely affecting social stability."

In March 2021, for instance, China's internet governance top bureau, the Cyberspace Administration of China(CAC) summoned 11 heavyweight Chinese internet companies for talks with the regulators, including Tencent, Xiaomi, Kuaishou, ByteDance, Alibaba, netease Music and Ximalay.[419] CAC and other governing bodies intend to conduct onsite inspection to push China's internet players to follow the relevant laws and regulations to regulate deepfakes. Though at this moment, there has been no litigations reported involving deepfakes, the Chinese central government has taken the negative social effects of deepfake rather seriously.

It responds to the negative impacts swiftly by making new rules on networked audio-video services and products, and updating related industrial standards for implementation purposes. Such measures are meant first to address the escalating concerns of disinformation and misinformation that threaten public order and national security as policy priority; and second to tackle the negative impacts on individuals' civil rights, including personal dignity, reputation, privacy and personal information/data protection. Further, China has issued a few key sectoral rules and policies to regulate networked audio-video information products, covering deepfake issues; but unlike the EU, at the moment China does not take, and intends to take, a systematic approach to regulate AI, even though AI and the related facial recognition technologies have deeply penetrated in the Chinese community for surveillance and commercial purposes. Further, deepfakes and the related issues can be regulated under China's traditional legal framework based on civil rights protection and remedies, as well as previous data privacy laws, to different extents. A last distinguished factor that needs equal

attention is the recent overhaul of China's civil law system with a new Civil Code valid as of Jan. 1st, 2021, which has fundamental impacts on civil rights protection in the long run as a game changer, e.g., the recognition of personality rights, including the rights to one's own image, reputation, honor, privacy and protection of personal data.

This country report only discusses the regulatory legal framework at the national level that addresses deepfakes and the related technological deployments. Given the fact that China has a very strong central government that is active in policy-law making in tech regulation and most provincial laws are in principle implementing national laws, it does not make sense to further discuss provincial legislations and rules in a country report like this. But one has to realize the fact that provincial legislations can exist to regulate the collection and processing of facial images and other personal information to protect individuals or to address related legal issues. For instance, the Social Credit Regulation of Tianjin Municipality has been regarded as the first local law to "forbid" collection of facial identification information. Invalidated on Jan. 1st 2021, Art. 16 forbids market credit organizations to collect personal information without their consent and lawful grounds, including natural person's religious belief, blood types, health condition and medical history, biometrical information, and other information that are not allowed for collection in China.[420] Further, there exist a rich set of industrial standards in network security and content regulation that are relevant for regulating deepfake technologies to implement state laws and administrative rules, however, this report only discusses the most relevant ones due to the limited space of a country report like this, while referencing to these documents in the report for reader's further references.

The report is structured as follows. Section 2 will provide a brief overview of China's deepfake technology regulatory framework. Section 3 discussed the traditional regulatory apparatus to address deepfake issues including the new Chinese Civil Code, Criminal law, data protection law, copyright law, AI regulation/policies, etc., while Section 4 will discuss more specific legislations and sectoral policies for deepfake regulation.

2. An overview of the Chinese regulatory framework

Based on the technological analysis of the deepfakes,[421] the current Chinese deepfake regulatory framework can be categorized into three major blocks of laws that regulate: a) video/visual based products or services concerning facial and body images, b) audio based products or services concerning human voices or audio information; and c) AI products or services processing video-audio data concerning human images and voices. Below is a brief overview of the Chinese laws and regulations that provide a framework to regulate deepfake technologies mainly in these three areas.

First, under the current Chinese legal system, any use, abuse and misuse of personal audio-video data/information will be subjected to China's data protection law in general, with penalties and remedies to curtail data infringement associated with deepfakes. Also the legal duties and obligations of data controllers and processors, or services and products providers in the Chinese context, in these data protection and privacy laws will regulate producers of deepfakes and align them with legal requirements accordingly from the data and privacy protection perspective.

Identified and identifiable facial (and body) images, the related video data, and human voice data that are involved in making deepfakes are defined as personal sensitive information (个人敏感信息), and personal biometric (identifying) information (more specifically) in Chinese law. For instance, the Personal Information Security Standards (GB/T 35273-2020), which is applicable to personal information processing activities carried out by all kinds of organizations according to the legislator, directly defines facial features(images) and other biometric data(explicitly incl. images and voice prints) as personal sensitive data for protection(Art. 5.4, Note 3, & Art. 6.3 (c)).[422]

Further, Art. 29 of China's draft Personal Information Protection Law (PIPL), which is a comprehensive data protection law made similar to the GDPR and currently in the final review, defines sensitive personal information as "Sensitive personal information means personal information that, once leaked or illegally used, may cause discrimination against individuals or grave harm to personal or property security, including information on race, ethnicity, religious beliefs, *individual biometric features*, medical health, financial accounts, individual location tracking, etc.", and thus considers biometric information as a category of sensitive data.[423] Personal information is also defined by the Chinese Cybersecurity Law (another key legislation) as all categories of information, recorded electronically or through other means, that can alone or together with other information, to sufficiently identify a natural person's identity, e.g., names, birth dates, ID numbers, personal biometric information, etc.[424] Once an individual's privacy and data protection rights are violated by deepfakes, they may request protection and remedies from Chinese court. For instance, a Chinese local scholar has won the first-ever facial recognition technology case at a local appeal court to delete his finger print and profile photos, when he refused to use the new facial image based entry administration system of the Hangzhou Safari Park that replaced the previous finger print based system, which rejected his entry without registration and activation.[425]

Second, there exists a body of substantial administrative policies and industrial standards to address network video-audio products and services (mostly mandatory, although with some degree of self-governing nature) that are categorized as network data security and content regulation issues at the national level. They are made by China's tech regulatory bodies responsible for industrial policies, content regulation, standardization, network security, national security, etc., including the Cyberspace Administration of China (CAC), State Administration for Market Supervision, Standardization of Administration, the Ministry of Culture and Tourism, and the National Radio and Television Administration, as well as other self-regulatory bodies, to implement laws in various sectoral areas. Such regulatory documents and acts cover deepfake issues in terms of content regulation (video-audio works), security standards, professional SOPs (Standard Operating Procedures) and technical standards. For instance, the National Standards of Distant Facial Recognition Systems (GB/T 38671-2020) directly requests tech instruments installed to detect and prevent the use of fake facial images. More importantly, the new Network Audio-video Information Services Regulation explicitly imposes specific obligations for audio-video

service providers to address deepfake technology deployed in disinformation by clearly identifying and marking out such information or contents, and explicitly forbids the use of deepfakes in production and distribution of news reports.[426]

Third, at an overarching level and somehow indirectly, deepfakes are also regulated by China's new Civil Code and Criminal Code in terms of torts and criminal penalties. China's new Civil Code recognizes the interests and rights in an individual's personality and dignity, protects a person's reputation, images, the like, privacy, and personal data, and provides tortuous remedies accordingly in a systematic way. In particular, when the abuse and misuse of such personal information (e.g., facial images, body images and voices used in deepfakes) will endanger public security and public order, and harm other's dignity and personality in serious ways, rights violators will encounter criminal penalties. For instance, China has, to the best, made it criminal offense to publish deepfake videos created by means of AI or VR,[427] in case there is no clear clarification of the nature of the related information. The use of deepfake to humiliate or discriminate against an ethnic group can be punished under Art. 250 of the Chinese Criminal Code with serious consequences. On the whole, China can rely on her traditional, formal legal framework to successfully address deepfake issues, given the recent promulgation of the new Civil Code and the updated Criminal law in a categorical manner as an overhaul to protect individuals fundamental rights for the purpose of legal reform. These traditional legal instruments also include China's copyright law, consumer protection law, AI regulatory policies (not hard law, but policies of soft law nature) etc. But their success largely depends on strict law implementation and compliance by means of the related technical and industrial rules and standards that set up proper baselines for law compliance and judicial criteria.[428]

In summary, there are five big blocks of law that are relevant for deepfakes regulation, in terms of legal authority and efficiency, including: a) national statutory laws protecting individual's fundamental rights to personality, dignity, image, reputation and autonomy i.e. the Chinese new Civil Code, criminal law, and Constitutional Law; b) data and privacy protection laws securing the control of individual data and private life; c) general regulations and rules regarding audio-video and AI services and products; d) copyright law protect the economic and personality interests of audio-video product creators; and e) other direct, specific administrative rules and industrial policies (and standards) directly addressing deepfakes. The related rules, regulations and policies will be further discussed in more detail in the following two sections in a more selective manner based on their relevance and significance to address deepfake related legal problems. In particular, not all tech and industrial rules and standards will be discussed, but the key ones that are directly relevant for deepfake regulation.

3. General state law and legislations

3.1 Fundamental rights protection: dignity, privacy and liberty

Art. 38 of the Chinese Constitution (amended in 2018) protects the dignity of Chinese citizens from insult, libel, false accusation or false accusation by any means. Art. 40 protects individual's privacy and freedom from invasions by any organizations or individuals, and their

freedom and privacy of correspondence. Art. 51 prescribes that in exercising freedoms and rights, individuals must not infringe upon the interests of the lawful freedoms and rights of other citizens in general.

3.2 The Chinese new Civil Code: personality rights and tort liability

China's new Civil Code, which is effective as of 1 Jan. 2021, has been the most important legislation in decades to systematically protect civil rights by adjusting to the new digital era. In general, personal images are projected under *personality rights*. Art. 109 protects personal liberty and dignity of natural persons; Article 110 the right to name, likeness, the right to reputation, the right to honor, and the right to privacy. In particular, Art. 111 prescribes that a natural person's personal information is protected by law; to access other's personal information must comply with state law and guarantee the safety of such information, and one may not illegally collect, use, process, or transmit other's personal information, or illegally trade, provide, or publicize such information. Art. 127 especially protects data and online virtual assets.

Art. 990 defines personality rights as the rights enjoyed by persons of the civil law, including the rights to life, to corporeal integrity, to health, name, likeness, reputation honor, privacy and the like, as well as other personality rights arising from personal liberty and dignity; and such rights may not be waived, transferred, or inherited (Art. 992), while the name, likeness, or the like of a person may be used by others upon authorization, unless not allowed by law or based on the nature of the right. Art. 955 prescribes that when exercising such rights, one can request the actor to stop the infringement, revoke the nuisance, etc., and the provisions on limitation periods shall not apply. Different from the EU data protection law, Art. 994 of the Chinese Civil Code protects the *personality rights of the deceased* including their names, likeness, reputation, honor, privacy or the like, and the spouse, children and the parents, and the relatives of the deceased have the right to request the actor to bear civil liability in accordance with the law. This means that in case of the abuse or misuse of the dead's images and voices in deepfakes, their relatives and offspring can gain legal remedy in litigation.

Art. 1010 protects persons from sexual harassment against his/her will by another person through oral works, written language, images, physical acts or the like. Further, Art. 1018 grants an individual the right to his likeness and the entitlement to make, use, publicize, or authorize others to use his own image according to Chinese law, further defining the likeness/personal images as "an external image of a specific natural person reflected in video recordings, sculptures, drawings, or on other media by which the person can be identified." Article 1019 (on rights to likeness, under personality rights) prescribes that "No organization or individual may infringe upon other's rights to likeness by vilifying or defacing the image thereof, or through other ways such as falsifying other's image by utilizing information technology. Unless otherwise provided by law, no one may make, use, or publicize the image of the right holder without the latter's consent." Without consent of the person holding the right to likeness, no one shall use or publicize the said image by publishing, duplicating, distributing, leasing or exhibiting.[429] The Article prescribes that the appropriation of the likeness as a tort does not need to be only for commercial purpose,

but includes usage of other's likeness, or ICT means for purposes of defamation, slur, or harm other's images or likenesses, which certainly includes deepfakes.

Article 1020 lists exceptions for the use of one's images, without consent, including: a) using publicly available images for personal study, art appreciation, classroom teaching, or scientific research; b) making using or publicizing individual images when inevitable for conducting news reporting; c) making, using or publicizing the images to the extent necessary for a state organ to perform its official responsibilities; d) Making, using, or publicizing the images for demonstrating a specific public environment; e) performing other acts of making, using or publicizing the images of a person for protecting public interest and lawful rights or interests of that person.

Article 1021 further protects individuals when a contractual dispute over the use of an image of a person authorizes the use by requesting interpretations made in favour of the person. Article 1023 protects an individual's likeness and voice in the context of authorized use by applying the *mutatis mutandis* doctrine. Further, in a more general line, Article 1024 (on rights to reputation as personality right) protects a person's image when there is harm against one's reputation that "A person of the civil law enjoys the right to reputation. No organization or individual may infringe upon other's right to reputation by insulation, defamation, or the like." Likely, Article 1025 in particular makes it accountable when a person's acts, i.e., reporting news or supervising public opinions or the like for public interest, adversely affects that personal reputation, when the like has been used to degrade the other's reputation." Article 1031 provides similar protection in terms of personal honour.

The new Civil Code also protects one's facial images under the rights to privacy and personal information in general, prohibiting collecting such information without consent and lawful grounds (Articles 1032-1039). Specifically, Article 1033 forbids taking photographs of another person or disclosing the private activities of the person (Clause 3), processing another's private information (Clause 5) and infringing upon another person's privacy via other means (Clause 6). Further, Article 1034 protects a natural person's personal information which is defined in a way to include one's biometric identifying information (Clause 2), such as one's facial image; and Article 1035 requests processing of personal information to comply with several principles of lawfulness, justification, purpose limitation, and to meet certain listed conditions (consent, transparency, purpose limitation and according to law, etc.), with three listed limitations to civil liability (Article 1036). Article 1038 forbids data controller's and processor's disclosure, or alteration of the collected or processed personal information, and provision of personal information to others *illegally* without consent of the data subject. This covers deepfakes as image alternation logically, and processing without explicit consent, unless such information/data cannot be used to identify a person.

Under tort liability, the new Civil Code has laid down detailed clauses for infringements against Chinese citizens by network users and service providers. Art. 1194 prescribes that internet users and service providers shall take civil responsibilities when they infringe other's civil rights by means of the internet. Art. 1195 clarifies

that in case network users use the internet to conduct such infringement against others, the latter has the right to notify the network service provider to take necessary measures, including deletion, blocking and unlinking. Such notice shall contain the primary evidence of the infringement and the real personal information of the infringed. Upon receiving the notice, network service providers shall transfer the notice to the network user and take necessary measures in due time according to the infringements; otherwise, the service provider shall take joint responsibilities. Art. 1196 requests internet users, upon receiving such infringement notice from the service provider, may submit a declaration of no infringement as accused, which shall contain primary evidence and the user's real identity information.

3.3 The Criminal Code: protecting personal reputation and dignity

The Chinese Criminal Code punishes perpetrators who may use deepfake technologies to cause serious infringement of reputation (of both individual and business) and dignity, serious discrimination against racial and ethnic groups, and production and dissemination of porn materials.[430] Art. 221 punishes the acts of fabricating stories and spreading them to damage another person's business credit or commodity reputation with heavy losses and serious circumstances. Art. 246 of the Chinese Criminal Code protects a person from being publicly humiliated, defamed publicly by violence or other methods with various punishments. Art. 363 punishes production, publication, sale, or dissemination of pornographic materials for the purpose of profit with imprisonment. This includes production or duplication of such pornographic audio-video products, the arrangement of such shows, or the dissemination of such materials. Art. 367 clarifies that this includes obscene books, periodicals, movies, video-and audio-tapes, pictures, etc. that explicitly portray sexual behaviour. Art. 246 lays out penalties for public humiliation and defamation of another person by violence or other methods when the circumstances are serious (but on the base of complaint only, except existing serious harm to public order or to the interest of the state). Further, Art. 250 punishes a person who publishes any content designed to discriminate or humiliate an ethnic group, if the circumstances are flagrant and the consequences are serious. Art. 253 punishes the invasion of private information in serious circumstances of which facial images can be protected under personal information in a wider interpretation as in Chinese law.

3.4 Data protection law

At the moment, China does not have a comprehensive data protection law like the GDPR. But a similar, more simplified legislation is under consultancy for final promulgation.*431* Besides this new Personal Information Protection Act (draft) (个人信息保护法（草案）), multiple formal legislations and state administrative policies and regulations can protect personal images and voices from being misused and abused in deepfakes.

First, the new Personal Information Protection Act (draft) was released for consultation in Oct. 2020. This new legislation is meant for general protection of individual information/ data to prevent their mis/abuse by enterprises, organizations and individuals for business purposes, leading to harms against individual's healthy, properterial and personality interests. Art. 13 sets up the conditions for lawful processing

of personal information, including consent and contractual obligations, other legal obligations, public health crises or other emergencies to protect natural persons' life and property, and news reports for public interests, etc. Art. 24 requests that in case personal information is provided to a third party, the controller shall inform data subject the third party's identity, contact, and the purposes, manners and categories of processed information, and the third party can only process within the notified scope; the third party shall notify the data subjects in case of any changes and regain consent for such processing. Art. 25 permits automated decision making concerning individuals on the condition that such decisions are transparent and fair in terms of consequences; the individual can request an explanation if he/she finds that this significantly impacts his/her interests and rights, and has the right to reject the decision, including business promotions, information distribution, etc. Further, Art. 26 forbids the disclosure of data subject's information by data controllers and processors to others, except based on consent, legal and regulatory grounds.

Art. 27 prescribes that installation of tech devices in public spaces for collection of human images and identification of individuals shall be necessary for the protection of public security, following related laws, and setting up warning signs for such devices. The collected images and identifier information can only be used for public security purposes, are forbidden to be publicized and provided to third parties, except after acquiring individual's consent or based on exceptions of state law, and administrative rules.[432] Art. 27's special regulation may prevent the collected personal images to be further used for illegal purposes such as deepfakes. Art. 28 requests that the processing of publicly accessible (individual) information shall be in accordance with their original purposes when published; any processing beyond the original scopes shall follow this law to acquire data subject's consent with notification. When the original purposes are not clear, processing shall be conducted with due care and properly; processing shall notify the data subject and acquire consent when there will be significant consequences on individuals. Art. 29 defines sensitive personal data and lays out more protection with stricter processing requirements.

Second, China's Cyber Security Law, as one of the key legislation in data protection and cybersecurity, defines personal information as all kinds of information, recorded electronically or through other means that are sufficient to identify a natural person's identity, alone or jointly with other information, including personal biometric information.[433] Art. 12 forbids individuals or organizations to use the internet to carry out activities to disseminate disinformation and fake news to interrupt social and economic order, to advocate racial discrimination and hatred, and to harm other's reputation, privacy, IP rights and other legal interests.

Last, among other data protection policies and rules, it is important to mention the Personal Information Security Standards (GB/T 35273-2020, 个人信息安全规范), which is made for the protection of data protection rights very similar to the GDPR standards. Validated on 1st Oct. 2020, this guideline is an important policy guideline of soft law nature setting up sectoral standards for the implementation of related state laws in a very detailed manner.[434] The proposed guidelines contain different standards regarding collecting biometric information (including personal

images, and facial images), their storage, sharing, and disclosure.[435] Art. 5.4 c) requires that before collecting personal biometric information, data subjects shall be *personally* informed of the collection, the purposes, manners and scopes of such collection, and the storage period, and such collection shall be conducted on the condition of explicit consent; and biometric information includes genetic information, finger print, palmprint, voice, iris, and facial features (Note 3). Art. 5.5 requests data controllers to set up internal policies to protect the collected personal information, not disclosing such information to third parties, clarifying potential risks, etc.

Art. 6.3 prescribes that biometric information shall be stored *separately* from personal information; and more importantly, personal biometric data (e.g., sample or images) in principle s*hall not be stored and kept,* and possible protective measures shall include: 1) only store the extracted information from personal biometric information, 2) directly using personal biometric identifiers to identify and authenticate individual in the data collection terminals, and 3) after using facial image traits, finger prints, palm image, and iris data for identification purposes, the original images that can be used to extract personal biometric information should be deleted. Further Art. 7.3 lists a number of circumstances (and purposes) in which biometric information cannot be processed, when such use shall not exceed the claimed purposes (direct and reasonably related scopes, otherwise, further consent is necessary).

In particular, Art. 7.4 limits the use of profiling, and requests data controllers: 1) not to use the description of data subject's personal traits in personal images to include: a) sexual, vulgar, gambling, violent, and terroristic content; and b) to express bias and disrespect against nationals, racial groups, religions, disabled people and other diseases; and 2) in their business operations or corporations with foreign parties the use of personal images shall not: a) violate the legal interests of other citizens, legal persons, and other organizations; and b) harm national security, honour and interests, promote racial hatred and violence, advocate porn or sexual information, disseminate disinformation, etc. These requirements will forbid personal facial images to be used in harmful ways against the listed interests.

Further, Art. 9.2 (i) in principle forbids sharing and transferring of biometric data, with the exception only when it is necessary for carrying out business and on the condition of notifying the individual data subjects of the purposes of processing, the categories of biometric information involved, the identity of the recipient(s) and its security capacity, and acquire the data subject's explicit consent. Art. 9.4 (f) does not allow public disclosure of personal biometric information.

3.5 Copyright law

In general, Art. 10(4) of the Chinese Copyright Law (amended in 2020) protects the integrity of one's works (e.g., personal images) from distortion, alternation, etc. Art. 10(1) protects the rights of the author regarding whether or not to publicize his/her works.

Art. 39 protects the images inherent in a performer's performance against distortion of personal images, publication of one's performance lively (incl. via the internet), and acquiring related income, authorizing the use of such images and sound recordings, and the reproduction (incl., through information

network). Art. 41 allows no time limit to protect the rights to claim performership of one's image inherent in one's performance, and protects such images from distortion, while allowing max. 50 years of protection of other rights of one's performance.[436]

Art. 24 introduces exceptions for the use of other's works without permission and payment, such as on the condition of indicating the author(s) and names of the used works, and on the condition the reuse won't influence the used work's normal usage; such exceptions also include, for instance, individual uses for study, research and entertainment; using other's works for the purposes of introduction and communication of a work, or explanation of certain problems; for news reports, journal publication, broadcasting purposes, etc.

Section 3 of the Copyright Law specially regulates the use of audio and video works. Art. 42 prescribes that when using other's works, audio-video product authors shall gain other's permission and pay related costs. But musician works (audio) can be used without permission in order to produce audio works with payment; unless the author forbids such use.

Art. 42 lists civil responsibilities(i.e., stop infringements, make apologies, rectify negative impacts, claim remedies, etc.) for alternation and distortion of other's works, etc. Art. 53 also lists serious infringements that lead to criminal responsibilities including: copying, disseminating, or publicizing via networks to the public audio-video products without author's permission, copying, and publicizing audio-video products containing the performer's performance without the performer's permission, or disseminating the performance via networks.

Creation of deepfakes involves the use of existing video or audio works of others and this may infringe copyrights of others protected by the Chinese Criminal Code. Art. 217 punishes the infringement of copyright for the purpose of making profits with illegal gains relatively large or causing serious circumstances, including reproducing and distributing motion pictures, or other visual works, or other works without permission of the copyright holder; and reproducing and distributing an audio or video recording produced by another personal without permission of the producer.

3.6 Audio-video Products Regulation

Art. 3 of the Chinese Audio-video Products Regulation forbids any contents in audio-video products in China that, among others, harms public security, sate honour and public interests, provoking racial hatreds, racial discrimination, and traditional costumers, endanger social stability and social order, promote vulgar and pornographic contents, defame and humiliate others and harm other's legal interests, and discriminate social morals, etc.[437]

3.7 Consumer Rights Protection Law

Article 29 requests that business operations collecting and using consumers' personal information shall abide by the principles of legality, rightness, and necessity, explicitly stating the purposes, means and scope for collecting or using information, and obtaining the consumers' consent. Proprietors must not collect or use information in violation of laws, regulations or agreements between the parties, and must disclose the data collection and processing rules. It also requests proprietors

and their employees must keep consumers' personal information strictly confidential, and not disclose, sell, or illegally provide consumer's information to others; they must take necessary technical and other measures to prevent consumer's data from being disclosed and lost.

Art 56 prescribes punishment for business operators that violate consumer's dignity, liberty and other rights in protection of consumer's personal information.

3.8 AI regulation

Artificial Intelligence (deep learning, machine learning, or automated data processing) has played a critical role in deepfake technology development. Deepfakes can be regulated in the EU by the recently drafted AI Regulation. China, however, has not made any formal legislation, or is intending to, regulate AI technologies in a systematic way, although AI has been one of the Chinese state's technological strategic focuses and China is considered one of the world' leading AI powers besides the US. To further enhance AI technology development has been one of China's national industrial priorities alongside digital economy, internet finance, and big data, cloud computing, following China's tech law legislation plan according to the CCP's Rule of Law strategy 2020-2025.[438]

On the whole, there have been a few key strategic industrial policy documents regarding AI development that outline China's AI national strategies and regulatory and ethical principles(binding, but not formal law). In specific, there has been no specific formal rules in these AI policies addressing deepfake issues.

A key document is the New Generation of Artificial Intelligence Development Plan which was issued by the Chinese State Council in 2017 (No. 35). It lays out China's AI development general plans, principles, aims and industrial focuses.[439] The Plan clarifies: the importance of developing laws, regulations and ethical norms to promote and ensure a healthy development of AI, including civil and criminal responsibility, law compliance, privacy and property protection, information security utilization associated with AI apps, in order to establish a traceability and accountability system.[440] It emphasizes on the establishment of moral standards and an intellectual property protection system for AI technology, by adhering to the principles of security, availability, interoperability, traceability, and gradually establishing and improving the basics of AI, interoperability, industry applications, network security, privacy protection and other technical standards.[441] Last, the Plan also outlines the safety supervision and evaluation mechanisms for AI development, by strengthening the research and evaluation of influences of AI in national security, and improving the security protection systems of humans and establishing an early warning system to monitor the safety of AI practice. [442]

More relevant is the 2019 Beijing AI Principles (BAIP) backed by the Ministry of Science and Technology (MST), which aligns with existing international AI principles, but also demonstrating distinctive Chinese characters such as a more practical approach to AI ethics, the underlying philosophy of optimising symbiosis (Chinese ethics of harmony rather than competition), and a more forwarding looking attitude.[443] Further it is worth mentioning the Governance Principles for a New Generation of AI backed by the Ministry

of Science and Technology (MST) in 2019.[444] This policy document combines most international AI regulatory standards and previous Chinese AI policies, characterized by its direct referencing to "human-machine harmony."[445] It states AI principles including fairness, respect for privacy, safety and security, joint accountability, and swift governance.[446] The principle of respect for privacy emphasizes on the significance of individual privacy protection, secures the right to information and individual free choice, and requests boundary making in personal information collection, storage, processing and deployment against misuse of personal information.

4. Special administrative rules and standards on audio-visual products

In China, deepfakes falls directly under content regulation from a regulatory approach. Multiple regulations, administrative rules and industry policies have been issued in this regard. They contain specific regulatory rules and technical standards for auditing and monitoring networked audio-video products in different formats. These legal documents regulate in a detailed way what are lawful contents and who are responsible for regulating network contents, as well as responsibilities for potential harms and offences. This includes:

- China's Network Information Service Regulatory Measures
(互联网信息服务管理办法);[447]
- Network News Information Service Administrative Rules
(互联网新闻信息服务管理规定);[448]
- Provisional Rules for Network Culture Administration
(互联网文化管理暂行规定); [449]
- Regulation on Internal Content Self-censoring for Network Culture Institutes
(网络文化经营单位内容自审管理办法);[450]
- Administrative Rules on Network Audio-Visual Programs and Services
(互联网视听节目服务管理规定);[451]
- Notice on Enforcing the Regulation of Network Audio-video Program Contents
(关于加强互联网视听节目内容管理的通知);[452]
- General Standards for Auditing Network Audio-video Programs
(网络视听节目内容审核通则);[453]
- Regulation on the Administration of Non-Adult Programs
(未成年人节目管理规定); [454]
- Administrative Notice Regarding Issues of Network Live Audio-Video Services
(关于加强网络视听节目直播服务管理有关问题的通知); [455]
- Administrative Rules Regarding Short Network Video Product Platforms
(网络短视频平台管理规范); [456]
- Specific Auditing Criteria for Network Short Video Products
(网络短视频内容审核标准细则).[457]

4.1 Administrative Rules regarding Short Network Video Product Platforms and Specific Auditing Criteria for Short Network Video Products

Two important content regulations of policy nature deserve our full attention. The Administrative Rules Regarding Short Network Audio-video Product Platforms, whose Art. 3 (2) requests network platforms of short video services shall carry out their duties to protect IP rights of authors, and shall not cut and edit all categories of broadcasting and video works like films, TV series, network movies, etc; it forbids circulation of user uploaded movies, TV series, network

movies and other contents, and circulating such contents from registered organizations. Art. 3(3) requests short video product platforms to follow the regulations of qualified news reports and not to circulate or disseminate user made political-societal short video products, especially such products made by product producers whose qualification of news reports is unverified.

The Specific Auditing Criteria for Network Short Video Products is issued by China Netcasting Services Association (a state supported industrial self-governing body) in 2019 to tackle widespread disinformation and misinformation on the internet. It explicitly forbids 100 categories of short video products containing contents that a) extract speeches and videos of Chinese political leaders to distort the original meanings, or use other technical means, i.e., motion pictures, to exaggerate or distort political leader's special attitudes and modes (Clause 12), b) misuse and abuse the images or the like of Chinese heroes (Clause 17), c) humiliate, mock and harm racial emotions via language, pictures or music (Clause 25), d) extract video clips intentionally from law enforcement activities to create the effect of excessive use of forces (Clause 43), and e) display and advocate pornographic contents (Section 16), etc. These clauses can be used to regulate and ban deepfake products if the listed contents are involved.

4.2 The Network Audio-video Information Services Regulation

The most important legislation that directly regulate deepfake is the Network Audio-video Information Services Regulation（网络音频视频信息服务管理规定）. This Regulation is a special legislation to regulate audio-video information and contents on the internet, based on China's Cybersecurity Law, Network Information Service Regulatory Measures, Network News Information Service Administrative Rules, Provisional Rules for Network Culture Administration and Administrative Rules on Network Audio-video Programs and Services. Non-compliance of the Regulation may lead to administrative sanctions, direct fines and criminal penalties in accordance with the above listed laws.

Effective as of Jan. 2020,[458] the Regulation is made by three government bodies, the Cyberspace Administration of China (CAC), the Ministry of Culture and Tourism (MCT), and the National Radio and Television Administration (NRTA) and is thus more of a policy nature. Different from other sectoral policies, it focuses on information services providers that produce, provide, publicize and disseminate video/audio products via the internet, apps and other networked platforms, in particular addressing new technologies such as deepfakes applied in video-audio areas.

First, the Regulation tackles deepfake directly. Art. 11(2) explicitly forbids the use of deepfake and virtual reality technologies to produce, disseminate and circulate disinformation/misinformation, and fake news both by audio-video service providers and their users. It prescribes that any reproduction or referencing to audio/video news or information shall only be conducted based on audio-visual news information that is made by the institutes that are allowed by related regulatory laws.

Second, the Regulation imposes key regulatory duties on audio-video service providers. Network audio-video service providers are imposed with regulatory duties to monitor, filter and regulate

such audio-video information/content in their services in general. Network audio-video service providers are defined as entities that produce, distribute and disseminate audio-video works via networked platforms such as the internet and apps under Art. 2. Art. 6 requests that audio-video service providers need to first *acquire operating certification or qualification* to provide such services (as a condition for entering business), depending on their service areas for which different standards may apply for such qualifications under general law or administrative law. This sets up a higher bar for stakeholders to enter the market, and state regulators can disqualify service providers when they do not fulfil the imposed regulatory duties under the Regulation and other laws.

Art. 10 requests the conduction of security and risk assessment (following related state law) by network audio-video service providers that use deep learning and virtual reality technologies to provide network video services that are of social media nature, carrying out societal campaigns, advocation and promotion. Art. 11 (1) explicitly requests audio-video service providers to mark *in a clear manner* the production, dissemination and circulation of *any untrue* video information/content made by means of deepfake, deep learning and virtual realities by such service providers, or other new technologies. Art. 11 (2) *forbids network audio-video information service providers to produce, publish and disseminate untrue news and information that are made by means of new technologies and new applications of deep learning, virtual reality, etc.* (thus forbidding deepfake and deep learning technologies to be used in news information production, publication and dissemination); any circulation of audio-video news information shall be conducted lawfully based on the audio-video news contents and products made by the organizations that operate within the scope of state regulations (i.e., licensed operators).

Art. 12 directly requests service providers to regulate audio-video information/content published on their platforms by different means, including a) installation of monitoring and identifying technologies, b) stopping transferring of illegal, or unlawful information, taking further measures of erasure, prevention of dissemination, keeping related data records, etc., and c) reporting such contents to network, cultural and broadcasting regulatory authorities. Once network service providers find any information or contents that are illegal under Art. 11 (1), they shall stop dissemination and transferring, and further dissemination or transferring is only allowed after such information has been distinctly marked.

Art. 13 requests that service providers, once finding that audio-video service users make, disseminate, or publish disinformation (rumours) by using fake pictures and audio products based on deep learning and virtual reality, they should take measures immediately to clarify the situation, and report to relative regulatory bodies for archiving purposes.

Third, the Regulation also prescribes concrete measures for audio-video information service providers to follow to fight deepfakes and other misinformation based on deep learning and virtual reality technologies. For instance, Art. 7 requests the establishment of internal procedures to clarify responsibilities of security and safety of their personnel, to set up registration systems for service users, procedures for information

disclosure, information security, internal training, protection of minors, IP protection, etc. In particular, Art. 8 requests the identification of users based on their institutional codes, ID, mobile numbers (real identity), etc.; without identification of a person, there is no service for publication of information for the unidentified service users. Art. 14 further requests clear contractual clauses to clarify the duties and rights of both audio-video information service providers and their service users, in particular users' obligation to abide by Chinese law and regulations; also information service providers have the authority to take necessary measures, for instance, to restrict, suspend, or stop services, send warning, archive data, or report to public authorities in charge.

To guarantee the implementation of the above rules, Art. 18 prescribes penalty measures following other laws and rules, and clarifies that this can involve administrative penalties and criminal responsibilities.

4.3 National standards on Distant Facial Recognition Systems

Among other regulatory standards,[459] another key policy document regarding facial recognition technology is the National Standards on Distant Facial Recognition Systems (*GB/T 38671-2020*). The Standards sets up directly the technical standards for distant FRT deployment and is promulgated from 1st Nov. 2020 by the Chinese Central Market Bureau and the Committee of National Standards. The Standard prescribes the basic requirements for functionalities, configurations and security safeguards in the information systems that use human facial images to conduct distant identification at the far end of internet services. The prescribed standards are applicable to the research, test and management of such information systems mentioned above and the tech/legal terms for its application (Art. 1). The Standards has established a complete set of industrial standards to achieve the mentioned ends to protect individuals and their data privacy rights, preventing misuse and abuse, and other related security risks.[460] Art. 1 defines the aims of the Act as: "stipulating the functions, performance, security requirements and security assurance requirements of an information system that adopts face recognition technology for remote identity authentication on the server side." Further, the Standard also defines and includes live body detection within the regulatory scope (client side, Art. 4.2.3; server side, Art. 4.3.1). The Standards lays out very specific technical standards to ensure the security of facial image data in both basic and advanced contexts of facial data processing, from the collection, to storage, transmission, processing, and to the last deletion.

More related to deepfakes, Art. 6.2.6.6 requests tech measures to detect and prevent the use of user-fabricated facial image data, not exclusive to: fabrication and copying, fabrication from photos, video-fabrication (capable to detect and prevent most fabrications from massac, replacing and reproduction of video images), *Deepfake of human faces* (i.e., made from singular or multiple facial images, or 3-D reproduction), etc. Art. 6.2.6.10 requests service providers to immediately send warning in case of the detection of the above misuse or abuse; Art. 8.1.1.1 prescribes compulsory security data set for addressing facial image fabrication.

4.4 Security Requirements for Biometric Information Protection

The Chinese National Information Security Standardization Committee (SAC/TC260) has made multiple technical standards to protect biometric data, in particular facial image information and genetic data, following well accepted ISO standards. The most relevant and important standardization document relevant to deepfake regulation is the Security Technical Requirements for Biometric Information Protection (信息技术安全技术-生物特征识别信息的保护要求) (draft, 2019).[461] Art. 3.1 defines biometric data as data concerning bio feature samples, biometric characteristics, biometric feature, biometric model, biometric property, biometric template, and biometric references; or the collection of the above. Specifically, it points to palm prints, finger prints, iris, faces, DNA, etc., and classifies facial images as biometric reference (Arts. 4.1 & 4.3). This document contains specific tech standards regarding: a) analysis of the biometric features and biometric systems' threats and resolutions, b) the security standards for combination of biometric references and identify references; c) system models for biometric systems for the purpose of storing and comparing biometric features; and d) guidelines for protection of personal information in processing biometric feature identification.

Another very recent draft standardization document (Apr. 2020) is the Data Security Standards for Facial Recognition Technologies (信息安全技术-人脸识别数据安全要求) relevant for deepfake technology regulation. It lays out the basic security requirements, execution measures and regulatory standards that are applicable to all business operations in the facial recognition business (Art. 1). Art. 4 lists three scenarios for processing facial images, namely, a) facial authentication, b) facial identification, and c) facial image analysis. Art. 5 sets up the basic security requirements in the above three scenarios, including: a) following other data security standards, b) data minimization, c) taking necessary measures for data security to protect data subjects, d) no facial images collection of natural persons without explicit consent, e) having the security capacity matching data processing capacities, f) carrying out facial image identification and authentication shall meet the conditions that: 1) other non-FRT measures are clearly inconvenient or efficient than FRT measures, and b) not implementing FRT tech to minors below 14 yrs, etc.

4.5 Technical Standards for Protecting Personal Financial Information

There are special industrial standards to protect individual facial images and other identifying personal information in industrial practices, which this country report cannot cover in more details for the limited spaces. For demonstration, a telling example is the National Committee of the Standardization of Finance Technologies issued a new national standard in Feb. 2020: The Technical Standards for Protecting Personal Financial Information JR/T 0171-2020). The new sectorial law defines biometric identifying information as the most sensitive data (at C 3 level) and requests all financial institutions in China not to authorize or commission any institutions that have no financial qualification to collect C3 level information; it requests that financial institutions, or the authorized entities accordingly, shall - when collecting personal biometric information, transferring via public networks, and storing C3 level information - use

encryption technology to protect the collected biometric information.[462]

The Standard classifies personal financial information into three categories according to their harms, risks and impacts consequent to unauthorized use or alterations. C3 level refers to the information that may identify a consumer and whose abuse/misuse may harm the property and information security of the consumer; for instance, information stored on bank card chips, security codes, logon codes, trade code, and biometric information. C2 level information refers to the information whose abuse/misuse only leads to normal harms to consumers' property and information security, including information regarding ID, Passports, authentication code, transaction record, KYC information (Know Your Customer), home addresses, etc. C 1 level information leads to the least harm, such as the opening date of a bank account, institution, etc.[463]

5. Concluding remarks

Like in the EU and US, China has encountered fast developments of networked technology, AI and other Audio-video technologies that can process collected individual facial and body images, voice data and other personal information to create alternated human images and audio-video works that can be untrue or misleading, if not all used for illegal purposes. China has taken a few key specific measures including industrial policies and standards to address these issues, such as the Network Audio-video Information Service Regulation, first for protecting public security and social order, and second for protecting individual's rights to their personality and dignity, as well as commercial interests. The most recent Personal Information Protection Action (daft) is a big overhaul in the same direction, although it may differ from the GDPR in terms of detailed rules and legislative ends, in addition to these existing content regulation and data protection laws. They can to a large extent be implemented to tackle the negative effects of deepfakes, besides China's many detailed, granular industrial standards in the fields. On the whole, one has to understand that the systematic recognition of the personality right, and the rights to liberty and dignity under the new Chinese Civil Code and the updated Chinese Criminal Code, has indeed provided a better legal frameworks to protect individuals' infringed rights by deepfakes, granting individuals the rights to systematic protection and remedies (at least in theory) after the two most important laws are updated to adjust to a digital, networked world. In combination with China's evolving data protection laws and industrial regulatory policies and standardization rules, a new Chinese legal framework has been developed gradually, capable of providing sufficient regulatory means to tackle deepfakes and other new technologies.

10.1.2 Verenigde Staten

Author: **Andrew Roberts**[464]

The Regulation of Deepfakes in the United States

Although there are relatively few reports in the news media of particular uses of deep fake technology, there has been a significant amount of commentary on what it is possible to create using this form of technology, the uses to which it might be put, and the societal effects of those uses, particularly on democratic processes. Footage of Barak Obama using an expletive to

describe Donald Trump produced using deep fake technology has been widely circulated, and used to warn of the dangers presented by Deepfake technology.[465] Further examples involving political figures in the United States include fake video of Donald Trump urging Belgium to withdraw from the Paris climate agreement,[466] and footage of Nancy Pelosi, the Speaker of the House of Representatives, which had been altered to give the appearance that she was slurring her words. Footage of Mark Zuckerberg apparently admitting that Facebook's aim is to manipulate and exploit its users, which was created using deepfake technology also attracted widespread media attention. [467]

Much of the media reporting in the United States on problems associated with Deepfakes, has focused on two particular uses of it. The first is the creation of sexually explicit material without the consent of the person depicted. There is now an acknowledgement that the non-consensual dissemination of sexually explicit images captured using conventional recording equipment - activity in which the person depicted actually engaged – is a serious problem.468 But it has been widely reported in the media, that the most common use – by far - to which Deepfake technology is put, is the (non-consensual) creation of sexually explicit material. [469]

The second is the production of the kind of material depicted politicians described in the opening paragraph. The use of this technology to influence political discourse and the outcome of elections for public office currently appears to be far less prevalent, than for the production of sexually explicit material. But the risks that Deepfakes pose to the proper functioning of democratic processes has been highlighted in the press, and recognised in Congress.[470] The Senate Intelligence Committee's conclusion that a Russian internet research agency - supported by the Russian government - engaged in disinformation activities in the period prior to the 2016 US Presidential elections that were intended both to undermine the public's faith in the US democratic process, and to influence the outcome of that election by damaging Hilary Clinton's reputation and campaign, has heightened concern about the future use of deepfake technology for such purposes.[471] It has been suggested by one member that 'the threat of election interference is perhaps the most menacing and urgent' of the threats posed by the use of deepfake technology.[472]

This recent experience of election interference might explain – at least in part – why the United States appears to be at the forefront of legislative attempts to regulate deep fake technology. Several states have enacted 'deep fake laws'. Writing largely before enactment of these laws, academic commentators had expressed mixed views as to whether the law that pre-dated these developments adequately addressed problematic uses of Deepfake technology. Chesney and Citron, suggested that while it might have been possible for a person who image had been used without their consent to bring an action against the creator in copyright and tort law, there would be significant obstacles.[473] They suggest that the process of transforming the person's image into a fake, for example, might, ground a fair use defence.[474] Further, with respect to the possibility of using the law of defamation[475] and copyright law to recover damages where Deepfake have been created using images of politicians, Ice points out that political figures will not own copyright in much of the footage

that is taken of them, and that the courts have set a high threshold for establishing liability where politicians seek to recover damages in defamation. Although they are doubtful as to the possibility of private law actions being successfully used to address problematic uses of Deepfake, Chesney and Citron expressed greater optimism that certain uses of Deepfake technology might attract criminal liability under existing legislation. They suggested, for example, that federal cyberstalking laws could be used to prosecute the use of a Deepfake to intimidate the person depicted, or where publication or dissemination could reasonably be expected to cause substantial emotional distress.[476]

Although most States have non-consensual pornography statutes, Gieseke points out that these take the harm to be a violation of privacy. Because many Deepfake use images that are in the public domain, and in the United States the broad position is that a person does not have privacy interests in information that they have made public, such laws would not criminalize the use of publicly available images to create sexually explicit Deepfakes.[477] Furthermore, she suggests the fact that Deepfakes do not depict acts that are not 'fully real', they would not be treated as privacy violations for the purposes of these non-consensual pornography statutes. [478]

It will be seen in what follows, that generally, the proposals that have navigated the legislative process to become law, have been relatively narrow in scope. Almost all are concerned with the two particular uses of Deepfake identified above – non-consensual creation and dissemination of sexually explicit material, and attempts to influence elections for public office. Conversely, proposals for more expansive legislation that would regulate a wider range of uses, have failed to pass. Part of the explanation for this might lie in the constraints that are imposed by the guarantee of freedom of expression provided by the First Amendment to the U.S. Constitution.

The Report begins with some consideration of the extent to which the First Amendment is likely to act as a constraint on attempts to regulate the creation and dissemination of Deepfakes. Part II will provide an overview of the various ways in which such legislation has defined the media that it regulates. It will be seen that a range of terms have been employed. While some pieces of legislation (and proposals for legislation) use the term 'deepfake' or 'deepfake video', others that do not. Rather they define the media that will be subject to regulation in terms that are sufficiently broad to encompass use of Deepfake technology. Parts III and IV provide accounts of legislation that has been enacted to regulate, respectively, the use of Deepfake technology to create material that is intended to influence the outcome of elections for public office, and to create and disclose, disseminate and publish sexually explicit material without the consent of the person who is depicted in it. Part III provides the details of legislation that passed in Texas and California that deals with the former – attempts to influence elections – in quite different ways. The Texan legislature creates a criminal offence, while its Californian counterpart has established a cause of action in private law that can be used by those depicted in Deepfakes to obtain injunctive relief and damages. Part IV provides the detail of legislation intended to address the problem of sexually explicit material, that has been enacted in 3 States – Virginia, California, and New York. Here too it will be seen that differing approaches have been taken, California

and New York establishing a statutory cause of action in private law, and Virginia attempting to deal with the problem using the criminal law.

Having set out in Part II and IV, the details of legislation that has passed and is now in force, Part V provides details of various Bills that failed to pass. These are of interest as they are examples of proposals that were more ambitious – in terms of the range of harms addressed – than legislation that has so far been passed. Particularly, notable are Bills introduced in the Massachusetts legislative assembly, and the U.S. Senate that would have imposed criminal liability on those who us Deepfakes to facilitate criminal or tortious conduct. This Part also provides a summary of the most comprehensive set of proposals relating to Deepfakes that have so far been produced. These proposals set out in a Congressional Bill introduced in 2019,[479] would have required those who create Deepfakes to incorporate a 'Digital Watermark' identifying the material as such, and visual and audio disclosures alerting those who access them to the fact that they contain manipulated video and or/or audio. It would have created a number of criminal offences for various intended uses of Deepfakes, including the depiction of persons in sexually explicit material with the intention to humiliate or harass; an intention to cause violence or physical harm, to incite armed or diplomatic conflict, to interfere with official proceedings including elections, and to commit fraud. In addition to criminal offences it also proposed a cause of action for those depicted in Deepfakes without their consent. To support on-going attempts to deal with the problematic use of Deepfakes, it made provision for the creation of a Deepfakes Task Force that would be responsible for advancing the efforts of the U.S. Government to combat the threat to national security posed by Deepfakes, and to undertake and co-ordinate research into developing technologies.

1. First Amendment Constraints on the Regulation of Deepfake

Legislative attempts to regulate the creation and dissemination of Deepfake in the United States need to be understood in light of the protection of free speech that is guaranteed by the First Amendment to the U.S. Constitution. This provides, among other things, that 'Congress shall make no law... abridging the freedom of speech.' Despite its wording, the protection it affords is not limited to speech. It extends to other forms of expression including musical lyrics, theatrical performances, pornography (provided it falls outside the category of 'obscene material'), satirical works, and non-verbal expression intended to communicate ideas - such as the wearing or display of symbols as an act of protest. Although there does not yet appear to have been any Constitutional challenge to legislation that regulates certain uses of Deepfake technology, Blitz suggests that Deepfakes are the kind of video and audio creations that have traditionally been considered in First Amendment jurisprudence, to be a form of expression.[480]

By their very nature, Deepfakes constitute misrepresentations, and most analysis of the extent to which the creation and dissemination of them might attract the protection provided by the First Amendment takes as its starting point the U.S. Supreme Court's decision in *United States v Alvarez*.[481] In issue in that hearing, was whether a law criminalizing false claims about receiving military medals or honours violated the Constitution's prohibition on laws that abridge

free speech. The plurality found that, save in a few historical categories of speech, there is no general exception to the First Amendment's protection for false statements. While false representations were said to be of low value and consequently, should not attract strong Constitutional protection, the Court suggested that this did not mean they were not protected at all. There are various lines of reasoning that underpin the protection of false representations. Their value in the 'marketplace of ideas', is said to lie in the likelihood that the process of rebutting them will tend to draw out the truth, and secure support for ideas expressed in public discourse that consequently come to be seen as more valuable.[482] Their protection in political discourse is said to be justified on the grounds that allowing the government to regulate false statements would set the institutions of government up as the arbiters of truth and falsehood, and confer on the state – through its prosecutorial bodies – the power to censor those whose views do not conform to the dominant ideology.[483]

The protection provided by the First Amendment does not extend to all falsehoods, however. False 'speech' that causes legally cognizable harms will not attract protection, and legislation that imposes criminal or civil liability in respect of such misrepresentations will not fall foul of the First Amendment. Accordingly, legislatures may provide a cause of action for the harm to reputation that is caused by defamatory statements, or punish speech that is intended to incite violence without violating the First Amendment. First Amendment protection does not extend to deceptive representations that would tend to undermine the integrity of public institutions and public confidence and trust in them, nor to false statements that are intended to defraud others out of property or other things of value, such as job opportunities. Legislation that imposes criminal liability for impersonating government officials or giving fabricated evidence in legal proceedings will not fall foul of the First Amendment. However, to fall within the exception to the general protection that it provides, it must demonstrated that there is a direct causal link between the restriction imposed and the harm that is to be prevented.

As previously stated, legislation regulating the creation and dissemination of Deepfake that has so far been enacted, does not yet appear to have been subject to challenge on First Amendment grounds. However, it is clear that any general prohibition on the use of the technology would not pass muster. It is only possible to say with any confidence that legislation that imposes criminal liability or provides a cause of action in respect of 'legally cognizable harms' – the kind of harms that the Supreme Court has previously determined will lift the protection provided by the First Amendment – that will survive such challenge. Such fate – successful challenge - is unlikely to befall regulation of uses of Deepfakes that will damage reputation, violate property rights, and erode confidence and trust in public institutions.[484]

Blitz has suggested that the nature of Deepfakes justifies extension of the currently recognised categories of harm that remove Constitutional protection from false representations.[485] As it becomes easier to create Deepfake, and more difficult to distinguish them from authentic audio and video recordings, the more difficult it will be to identify reliable sources of knowledge. This, he argues, will undermine what he refers to as the 'knowledge ecosystem' that public discourse relies upon. The First Amendment should not give a person free rein:

> '[to] disguise the source of their claims in authoritative clothing by using technology such as video- or audio- fabrication or digital forgery to give it an appearance of reality that the expressive content alone cannot create. In other words, while the First Amendment gives someone the right in a discussion of public affairs to provide any answer they like – even a false one – t the question, "What should I believe?" and even the follow-up question, "Why should I believe that?", it doesn't give them the right to answer the latter question by creating an illusion rather than an explanation.'[486]

The First Amendment imposes a significant Constitutional constraint on attempts to regulate the use of Deepfake technology. Generally, the more narrowly focused attempts to regulate the use of Deepfakes are, the more clearly they identify the harms to be addressed, and the extent to which those harms align with harm that the U.S. Supreme Court has previously found to take expression ouside the protection afforded by the First Amendment, the more likely it will be to withstand challenge on Constitutional grounds.

2. Defining the Medium

In any attempt to pass legislation intended to regulate the creation and dissemination of Deepfakes, one of the initial challenges will be finding a definition that encompasses both the technology and techniques that are currently being used, and those that are likely to emerge in the future. It should be noted that the definitions discussed in this part of the report are found in legislation (or in Bills that failed to pass) that address harmful consequences of differing uses of Deepfakes. Where they are used in an attempt to influence those voting in elections for public office, the concern will be manipulation of voters. One way of ensuring that a cause of action or criminal offence is appropriately circumscribed – so that it does not criminalise or impose liability where use of the technology is unlikely to cause this form of harm (and ensure consistency with the protections provided by 1st Amendment) – will be to define the technology, in part, in terms of its capacity to deceive. In legislation intended to deal with the problem of non-consensual sexually explicit material that is created without the consent of the person depicted, the relevant direct harm will be its effect on the person whose image is used,[487] rather than any belief it might cause in the person who views it. But it is possible to abstract from various definitions that include references to substance of the material and the effects that it might have, the terms that have been used to describe the medium. This is the aim of this Part of the Report. However, because in later Parts there will be references to the definitions set out in this Part, what follows is organised according to the harm addressed by the legislation or proposals in which the various definitions are found.

(i) Influencing Political Elections

The first state to enact legislation to address the problem of Deepfakes being used to influence voting in elections for public office was Texas.[488] It criminalizes the creation and publication of material that is intended to influence the outcome of an election. As will be seen, the Texan legislation is unusual insofar as it uses the term 'Deepfake video' in the offence it created. That term is defined in a new paragraph that has been inserted into the Electoral Code. This provides that '"*Deepfake video*" *means a video... that appears to depict a real person performing an action that did not occur in reality*'.[489]

This definition is potentially problematic. Take the example of a Deepfake that is created by superimposing images of the head or face of person A, onto footage of person B performing the depicted act. The 'Deepfake video' depicts A performing an action that did occur in reality, albeit by person B. For the provisions to fulfil their intended purpose, words need to be read into the definition to make clear that a 'Deepfake video' is a video that depicts person A performing actions they did not perform. This issue is avoided in in Californian legislation that was passed around the same time as the Texan legislation, and which is intended to deal with the same problem.

The definition in the Californian legislation refers to *'video recordings'* but also to *'images'* (and audio recordings). It is therefore broader in scope than the Texan legislation. It also differs in the way that it describes the substance or nature of the altered material. While the Texan legislation refers to video recordings depicting *'acts that did not in reality occur'*, the material to which the Californian law applies is *'an image or an audio or video recording of a candidate's appearance, speech, or conduct that has been intentionally manipulated'* in a manner would cause certain effects that are related to the harm at which the legislation is directed. The term 'image' is so broad in scope that it could be construed to encompass any visual medium including material that is produced using Deepfake technology.

The legislation applies only to material that would (i) cause a reasonable person to believe that the material was authentic, and additionally (ii) to cause such a person 'to have a fundamentally different understanding or impression of the expressive content of the image or audio or video recording than that person would have of the person were hearing or seeing the unaltered, original version of the image or audio or video recording.' While the Texan legislation requires those applying it to resolve a seemingly difficult philosophical question – 'whether and what in an altered video recording represents what occurred 'in reality' - application of the Californian involves empirical questions that are perhaps more easily addressed – 'has the material has been manipulated?' and 'what effect is the manipulation likely to have on a person who views the material?'

(ii) Sexually Explicit Material

Three States have so far enacted legislation that addresses the problem of non-consensual creation and dissemination of sexually explicit material – California,[490] Virginia,[491] and New York.[492]

Virginia: The Virginian legislation applies to *any video graphic or still image created 'by any means whatsoever' that depicts 'another person'* (a person other than the person who created the video graphic material) who is totally nude or in state of undress that exposes genitals, buttocks or breasts. It is explained that the 'other person' referred to is one whose image is used to *'create, adapt or modify the videographic or still image.'* The term *'videographic'* is not defined in the legislative provisions.

California: As will be explained in more detail below, the Californian legislation creates a cause of action for those who are depicted in sexually explicit material. While the Californian law addressing the use of Deepfake technology to influence voting - referred to in the previous section - uses the terms *'video recordings'*, *'audio recordings'* and *'images'*, the legislation

concerned with the production of non-consensual pornography refers to *altered 'digitized' depictions*. The cause of action is available to persons who *'appear as a result of digitization to be giving a performance they did not actually perform, or to be performing in an "altered depiction"'*. It is explained that *'Digitization'* means to realistically depict (i) nude body parts of another person (or computer-generated nude body parts) as body parts of the person who will have the cause of action, or (ii) the individual engaging in sexual conduct in which they did not engage. The term **'altered depiction'** encompasses circumstances in which images of a person engaged in one form of sexual activity has been altered to make it appear that they are engaging in a different form of sexual activity. It is not clear why the two pieces of legislation use differing definitions.

New York: At the time of writing, the most recently enacted legislation regulating the use of Deepfakes – had been passed by the State of New York (in force April, 2021).[493] That legislation amends the New York Civil Rights Law to create causes of action in respect of the creation of non-consensual sexually explicit material, and unauthorized commercial use of a Deepfake that uses images of a deceased performer. The provisions that address the former issue, define the material concerned in terms that almost identical to those found in the Californian legislation that addresses the problem of non-consensual creation and dissemination of sexually explicit material.

(i) Unauthorised Commercial Use in Artistic Material

The legislation passed by the New York referred to in the previous section provide a cause of action for unauthorised commercial use of Deepfake created using that a deceased performer's image, includes a definition of that by comparison with that found in the provisions that relate to non-consensual sexually explicit material, appears sophisticated. A cause of action is established in respect of the unauthorised use of a *'digital replica'*. This is defined as:

> 'a newly created, original, computer-generated electronic performance by an individual in a separate and newly expressive sound recording or audiovisual work in which the individual did not actually perform, that is so realistic that a reasonable observer would believe it is performed by the individual being portrayed and no other individual.'

(i) A Generic Definition

The ground that has been covered in this section so far, reveals that legislation regulating the use of Deepfake uses a broad range of terms to describe the medium. This is, perhaps, an artefact of a fragmented legislative approach to harmful uses of Deepfake technology. It might be, in part, a result of the terminology commonly employed in the domain being regulated - the terms that tend to be used to describe material in that context. But there seems to be no reason why, whatever the harm that legislation regulating the use of Deepfake is intended to address, some common terminology could not be employed. The case for this– in terms of coherence and the accessibility of the law – is particularly compelling where the problematic use of Deepfake is addressed in various bodies of law enacted in the same jurisdiction, or a single piece of legislation that is intended to deal with various problematic uses of Deepfake technology.

An example of the latter can be found in a Bill introduced in Congress in 2019 – Congressional Bill HR 3230.[494] This Bill, the substantive content of which is discussed in Part VI of the Report, set out proposals for a range of criminal offences related to certain uses of an *'advanced technological false personification record'* that does not include a 'digital watermark' that enables it to be identified as a Deepfake, and a disclosure that informs viewers that the visual content has been altered. The offences concern the use of sexual images to harass or humiliate the person depicted, to cause violence, to incite armed or diplomatic conflict, commit fraud, and to influence an election. The definition of a Deepfake set out in the Bill is perhaps the most sophisticated of any that can be found legislation enacted to date. It defines an '*advanced technological false personification record*' as:

> *'Any deep fake which –*
>
> *(A) a reasonable person, having considered the audio or visual qualities of the record and the nature of the distribution channel in which the record appear, would believe accurately exhibits –*
>
> *(i) any material activity of a living person which such living person did not in fact undertake; or*
>
> *(ii) any material activity of a deceased person which such deceased person did not in fact undertake, and the exhibition of which is substantially likely to further a criminal act of result in improper inference in an official proceeding, policy debate, or election; and*
>
> *(B) was produced without the consent of such living person, or in the case of a deceased person, such person or heirs thereof.*

It goes on to explain that 'material activity' is 'any falsified speech, conduct or depiction which causes, or a reasonable person would recognise has a tendency to cause perceptible individual or societal harm, including misrepresentation, reputational damage, embarrassment, harassment, financial losses, the incitement of violence, the alteration of a public policy debate of election, or the furtherance of any unlawful act.'

'Deepfake' is defined as 'any video recording, motion-picture film, sound recording electronic image, or photograph, or any technological representation of speech or conduct substantially derived thereof –

> *(A) which appears to authentically depict any speech or conduct of a person who did not in fact engage in such speech or conduct; and*
>
> *(B) the production of which was substantially dependent upon technical means, rather than the ability of another person to physically or verbally impersonate such person.*

These definitions are found in statutory (or draft statutory) provisions that are intended to address at various problematic uses of Deepfakes, and it is to the substance of these that we now turn.

3. Influencing Elections

Legislation intended to address the use of Deepfake to influence the outcome of elections has been passed in 2 States – Texas and California. Differing approaches have been taken, one State creating a criminal offence, and the other establishing a cause of action in private law that can be used by candidates

whose campaigns have been targeted using Deepfakes. As indicated in Part I of this report, the First Amendment to the U.S. Constitution provides particularly strong protection for 'political speech', or put differently, to expression that makes a contribution to political discourse, even if it is false. However, legislation is likely to withstand challenge on First Amendment grounds, if it is narrow in scope, directed to a clearly identified harm, and a causal connection can be demonstrated between the restrictions that have been put in place and the relevant harm. These are characteristics of the legislation that has been enacted in Texas and California.

(a) Texas

Texas was the first States to pass legislation intended to deal with the use of Deepfakes to intended to influence the outcome of elections for public office. The Bill's sponsor explained that the proposal was motivated by a concern that "[a]s this technology continues to advance, it will become increasingly difficult to trust anything in public discourse." It was noted that while "[t]his technology likely cannot be constitutionally banned altogether, but it can be narrowly limited to avoid what may be its greatest potential threat: the electoral process.[495]

To this end, the legislation amends the State's Election Code, establishing a new criminal offence of (i) creating a deep fake video and (ii) causing it to be published within 30 days of an election, (iii) with the intention of injuring an candidate or of influencing the result of an election. The legislation provides no specific defences, and is a Class A Misdemeanour, the most serious of the lower level offences (Felonies being more serious). A person convicted of the offence can be fined up to $4000 and/or imprisoned for a period of 12 months.

(b) California

The State of California has also enacted legislation that is intended to deal with the problem of deceptive audio- visual-recordings that are created and deployed with the intention of influencing elections. However, the approach it has adopted differs from that taken by the Texas legislature. Rather than attempting to deter such conduct through the threat of criminal sanction, it has established a statutory private law cause of action that can be used by candidates in elections who have been targeted using manipulated audio- or video-recordings.

The cause of action is set out in Section 20010 of the Californian Elections Code,[496] is available to available (i) to a candidate for elective office, and (ii) can be commenced against a person (or political party committee) who (iii) within 60 days of an election at which the candidate appears on the ballot, (iv) distributes with actual malice, (v) 'materially deceptive audio or visual media' (see Part I of this report for discussion of this definition), (vi) with intent to injure the candidates reputation, or to deceive a voter into voting against the candidate.

Liability is limited in various ways. First, there will be no liability if the audio or visual media includes a disclosure stating that the image, or audio, or video has been manipulated. The following are exempt from liability for re-publication of the deceptive material: (i) broadcast news reports or documentaries provided they clearly state the media is deceptive, (ii) broadcast media that is paid to broadcast the materially deceptive

material, and (iii) internet sites that regularly publish news-content provided they clearly state that the material content does not accurately represent the speech or conduct of the candidate.

Those who succeed in establishing liability may be granted injunctive or other equitable relief prohibiting distribution of the audio or visual media, and can bring action for specific or general damages against the person, committee, or other entity that distributed the deceptive audio visual material.

(c) U.S. Senate – Bill in Progress

When this report was drafted, a Bill had been introduced in the U.S. Congress, which would - if passed - create both a cause of action and a criminal offence in respect of the distribution of the material that establishes a cause of action under the Californian legislation described in the previous section. The conduct that gives rise to liability is described in terms identical to those used in Section 20010 of the Californian Election Code. In the Congressional Bill,[497] that conduct is cast as a general prohibition, violation of which can either constitute the basis of a civil action by the candidate, or of a criminal prosecution. Civil liability arises in precisely the same circumstances as it does under the Californian legislation. Criminal liability depends on whether there is 'wilful and knowing' violation of the general prohibition. These convicted could be fined up to $100,000 and/or sentenced to a maximum of 5 years imprisonment.

4. Non-Consensual Sexually Explicit Images

Three States have enacted legislation that address the problem of the non-consensual creation and dissemination of sexually explicit material –Virginia, California, and New York. While Virginia has created a criminal offence of unlawful dissemination or sale of images, New York and California have chosen to establish a cause of action in relation to private law.

(a) Virginia

Virginia was the first to legislate, creating new offences that are found in Chapter 8 of the Code of Virginia – 'Crimes Involving Morals and Decency'. These are not specific to Deepfakes, but extend to the production of material using this technology and to its publication or dissemination. The first of the offences concerns the creation of sexually explicit material. Section 18.2.386.1 makes it unlawful for any person to (i) knowingly and intentionally create (ii) any videographic or still image by any means whatsoever (iii) of any non-consenting person (iv) if that person is totally nude, clad in undergarments, or in a state of undress so as to expose the genitals, pubic area, buttocks or female breast in a restroom, dressing room, locker room, hotel room, motel room, tanning bed, tanning booth, bedroom or other location.[498] The offence encompasses use of the kind of equipment and devices that will often have been used in the past to create sexually explicit images of persons who have not consented to their creation – mobile phones, and other video-recording devices. But in specifying that the offence will be committed by a person who creates a videographic or still image 'by any means whatsoever', the prohibition is stated in terms that appear sufficiently broad to impose criminal liability on those who use Deepfake technology to create proscribed images. However, if the words 'by any means whatsoever'

are read in the context of the provision taken as a whole, and in light of the terms in which the offence is described in the Code – 'Unlawful Filming, Videotaping, or Photographing of Another' – there are grounds for doubting that the provision will be construed as one that criminalises the production of Deepfakes. Further support for such a view can be derived from the explicit reference to the 'adaption or modification of a videographic' in the offence of unlawful dissemination of sexual videographic found in Section 18.2.386.2 of the Code.

Section 18.2.386.2 imposes criminal liability on those who disseminate or sell any videographic or still image - created 'by any means whatsoever' - of a person who is totally nude, or in a state of undress so as to expose the genitals, pubic area, buttocks, or female breast, if they know or ought to have known that they are not authorised or licensed to disseminate or sell the material.[499] The provision goes on to explain that 'for purposes of this subsection, "another person" includes a person whose image was used in creating, adapting, or modifying a videographic or still image with the intent to depict an actual person and who is recognizable as an actual person by the person's face, likeness, or other distinguishing characteristic.'[500] The implication of the inclusion this in the dissemination offence - and its absence from the provision criminalising creation of images - is that the term 'another person' in Section 18.2.386.1, does not extend to the modification of existing images. If this interpretation is correct, the Virginian legislation criminalises dissemination or sale of non-consensual sexually explicit Deepfakes, but not the creation of them.

The reason for limiting criminal liability in this way, it is not obvious. The harm caused by the use of Deepfakes to create non-consensual pornography, is usually taken to be the effect that such material is likely to have on the person depicted - the distress that publication will cause and forms of indirect harm that might flow from publication, the withdrawal or withholding of employment offers, for example. The mere creation of Deepfakes that depict a person in nude or engaged in sexual activity absent any disclosure, will not often give rise to this form of harm. The person depicted will usually – but not always - be unaware that they have been created. The harm will generally be a consequence of their publication. But this does not explain why the mere creation of sexual images using conventional video-recording devices ought to be the subject of criminal sanction while the creation of such images using Deepfake technology is not. Perhaps the reason is that where a recording device – typically a mobile phone – is used to create such images, the person who appears in them will be present and often aware of what has happened. But this will not be the case where recording devices are used covertly, in motel bedrooms, changing rooms, and for 'up-skirting', all circumstances that fall within the scope of the offence of unlawful filming in Section 18.2.386.1.

The offence of disseminating in Section 18.2.386.2 is committed only by those who maliciously disseminate, or sell with the intention to coerce, harass or intimidate. The significance of these mens rea requirements, is that platform owners and administrators who permit, or enable those who have created sexually explicit material - as defined in the provision - to publish it without the consent of the person depicted, will not be

liable. Section 18.2.386.2(B) removes any doubt by providing an exemption from liability for internet service providers, electronic mail service providers, and those who operate any other information service, system, or access software that provides or enables computer access by multiple users.

Offences committed under these provisions are Class 1 Misdemeanours, the maximum penalty for which is a fine of $2500 or 12 months' imprisonment, or both.

(ii) California

Whereas Virginia has criminalised the non-consensual creation and dissemination of sexually explicit Deepfakes, California has enacted legislation that establishes a statutory cause of action in private law for persons who are depicted in them. Section 1708.86 of the Californian Civil Code provides that a person who appears as a result of 'digitisation', to be giving a performance that they did not actually perform, or to be performing in an 'altered depiction' has a cause of action against (i) those who both create and intentionally disclose sexually explicit material, and who know or should reasonably have known that the person depicted in the material, did not consent to its creation or disclosure; and (ii) anyone who intentionally discloses sexually explicit material knowing that it has been created without the consent of the person depicted in it.

'Sexually explicit material' is defined in the legislation to include any portion of an audio-visual work that shows the depicted person performing in the nude, 'or appearing to engage in, or being subjected to, subject to "sexual conduct"'. The provisions define what constitutes both nudity,[501] and 'sexual conduct'. [502]

Because the liability of person who disseminates but does not create the material depends on them 'knowing' (as opposed to having reasonable grounds to suspect) as is the case with the Virginian criminal legislation, those who merely publish or facilitate the publication of it, are unlikely to be vulnerable to an action under this legislation. In most circumstances, those who own or administer platforms on which such material is published, or who otherwise disclose it, will be able to avoid liability by simply declining to make any inquiry to establish that the person depicted in the Deepfake has given consent to its creation. In any case, the provisions provide a number of specific defences. Those who disclose such material for the purposes of reporting unlawful activity, in the exercise of a lawful duty, or in the course of a trial or other legal proceedings, will not be liable.[503] Nor will there be any liability where the material is (i) a matter of legitimate public concern or (ii) has political or newsworthy value.[504] Perhaps to save the legislation itself from Constitutional challenge – as distinct from particular decisions about its applicability in any particular circumstances – the drafters make it clear that no cause of action exists in respect of any commentary, criticism or disclosure that is otherwise protected by the California or US Constitutions. Liability is also limited by a 3 year limitation period which runs from the point at which the material is discovered by the person depicted – or ought to have been discovered with the exercise of due diligence.

The provisions impose strong conditions on those seeking to create and disclose sexually explicit audio-visual work with the consent of the

person who appears in it. Consent is defined 'an agreement written in plain language signed knowingly and voluntarily' by the person who is depicted in it, and it will be valid only if given to the person 3 days prior to signing so that the terms of the consent can be reviewed. By making provision for the rescission of consent (which must also be in writing) within 3 business days of consent being given, the legislation also provides a 'cooling off' period for those who have given consent.

It is significant that the legislation makes clear that the inclusion of a disclaimer in the material - to the effect that it was not unauthorized by the person who appears in it, or that they did not participate in the creation or development of it - does not constitute a defence.[505] This suggests that the statutory cause of action is intended to address a range of harms, not just the distress that publication or other dissemination of sexually explicit material that might be perceived by those who view it to be authentic, could cause the person who is depicted in it. The range of remedies that are available tends to supports this view. Those who succeed in establishing liability can recover (i) damages to an amount equal to any monetary gain made by the defendant from the creation, development, or disclosure of the sexually explicit material, and (ii) economic and non-economic damages caused by disclosure, including emotional distress. The legislation also makes provision for the award of statutory damages for any unauthorised act with respect to any one work of between $1,500-$30,000. Where an unlawful act has been carried out with 'malice' – that is, with intent to cause the plaintiff harm, or 'despicable conduct done wilfully with disregard to the rights of the plaintiff' – damages of up to $150,000 may be awarded. In addition, courts are permitted to grant any other available relief including injunction.

Notably, the legislation omits a seemingly significant provision included in a Bill that would have criminalised the conduct with which it is concerned, which was introduced in February 2019 but failed to pass.[506] That provision would have permitted the plaintiff to proceed using either a pseudonym, and to redact from all pleadings and filed documents, any information or characteristics that might identify them, including name, address, relationship to the defendant, race or ethnic background, telephone number, email addresses, social media profiles, and other online identifiers.[507] It can be assumed that this provision was included in the Bill to address the distinct possibility that those whose image had been used in a sexually explicit Deepfake without their consent might be reticent to commence legal proceedings that would generate publicity regarding the existence of the material, drawing it to the attention of those who might not otherwise have discovered it, and increasing the likelihood of copies being made, and it being widely re-published.

(iii) New York

The most recent statutory regulation of the use of Deepfake technology for the purposes currently under consideration – in force from April 2021 – is that enacted by the State of New York. A Bill introduced in the New York legislature in May 2019, amends the New York Civil Rights Law[508] to provide a new statutory private cause of action that largely reflects that passed by the Californian legislature (described in the previous section). Many of the provisions are drafted in terms that are identical to those found in the

cause of action in the corresponding provisions of the Californian Code.

The New York provision differ from the Californian in only in a small number of respects. Sexually explicit material is defined by reference to pre-existing definitions set out in New York's Penal Code,[509] which are broader in scope than the sexual activity to which the Californian legislation applies. The limitation period for actions is potentially shorter under the New York legislation. In New York, an action must be commenced within 3 years of dissemination or publication, or 1 year from the date on which the material was discovered by the depicted person, or reasonably should have been discovered, whichever is the later. In circumstances in which the relevant date for the purpose of determining the limitation period is that on which disclosure/ publication etc *should reasonably have been discovered* (New York) or *ought to have been discovered* (California), an action in New York must be commenced within 1 year, in California the period is 3 years.

5. Unauthorised Use of Deceased Performers' Images

The only legislation to have been enacted which regulates the use of Deep Fakes for purposes other than attempts to influence the outcome of elections, or to create non-consensual sexually explicit material, concerns unauthorised use of a deceased person's image. The relevant provisions were passed as part of the Bill that also included provisions dealing with the creation and dissemination of non-consensual sexually explicit material described in the previous section. The Bill further amended the New York Civil Rights Law, inserting a new Section 50-f, which establishes a 'right to publicity' in respect of Deepfake depicting a 'deceased performer'. The provisions explain that a 'deceased performer' is a person who was domiciled in New York when they died, and who was regularly engaged in acting, singing, dancing, or playing a musical instrument, for gain or livelihood.[510]

The section provides remedies where there has been unauthorised use of a deceased performer's image in an artistic work producing using Deepfake technology.[511] This is done by enabling an action for damages to be commenced against anyone who (i) 'uses a deceased performer's digital replica in a scripted audiovisual work as a fictional character or for the live performance of a musical work... without prior consent', (ii) where the use is 'likely to deceive the public into thinking it was authorised.'[512] It is explained in the provisions that the right to publicity is a property right that is 'freely transferable or descendible, in whole or in part, by contract, gift, or by means of any trust, or any other testamentary instrument.'[513] Damages can be sought by any person who holds the rights to the material that has been used to create the Deepfake, for any injury caused by its unauthorised use. These can include any profits that are attributable to the unauthorised use, as well as punitive damages. The cause of action ceases 40 years after the death of the deceased person.[514]

There are a number of circumstances in which unauthorised used will not give rise to liability. The first is where the person using the image includes in the digital replica, 'a conspicuous disclaimer in the credits of the scripted audiovisual work, and any related advertisement in which the digital replica appears, stating that the use of the digital replica has not been authorised.'[515]

There are also exemptions for the use of digital replicas in works of parody, satire, criticism and commentary, and; documentaries, historical and biographical works.[516] Other circumstances in which those using a digital replica will not be liable include news reporting, public affairs and sports programming.[517]

6. Bills that Failed to Pass

In addition to the legislation that is now in force, a number of Bills have been introduced in the US House of Congress and State legislatures, but failed to pass. Some of these were precursors to bills that were re-introduced and subsequently passed, and their content did not differ substantially.[518] However, some contained proposals that differ substantially from any that can be found in the legislation that has so far been enacted. This part of the Report begins by describing proposals for a Californian Bill that proposed the a criminal offence of creating and disseminating non-consensual sexually explicit material. This is followed by a brief discussion of proposals to criminalise the use of Deepfake with the intention of facilitating crimes or tortious conduct. The final section describes what is perhaps the most comprehensive set of proposals to regulate the use of Deepfake that have so far been drafted – those set out in Bill H.R. 3230, which was introduced in Congress during 2019.

(i) Sexually Explicit Material

Although California and New York have passed legislation that provides a cause of action for those who appear in sexually explicit material created using Deepfake technology and subsequently disclosed, none that imposes criminal liability for such conduct has yet been passed. This would not have been the case had a Californian Bill introduced in February 2019,[519] successfully navigated the legislative process. This would have amended the Californian Penal Code to crearte a number of new offences. The first would have imposed criminal liability in respect of knowingly preparing, producing or developing a Deepfake that depicts a individual 'personally engaging in sexual conduct' *and* distributing or exhibiting it, or exchanging it with others. Those who had knowingly preparing, producing or developing a deepfake that depicts a individual under 18 years of age 'personally engaging in sexual conduct' and distributing or exhibiting it, or exchanging it with others, would also have committed a criminal offence that would have been created by the Bill. The maximum penalty for the former would have been a fine of $1000 and/or 12 months imprisonment, and for the latter $10,000 and/or 12 months imprisonment. The Bill also contained provisions that would have secured from Government funds, $25m for the University of California for research to identify and combat the inappropriate use of Deepfake technology.

(ii) Facilitating Criminal or Tortious Conduct

There have been two unsuccessful attempts to pass legislation that would, in practice, have imposed broad criminal liability. Bills were introduced in Massachusetts[520] and the U.S. Senate[521] that would have made it an offence to create and/or distribute a Deepfake with the intention of facilitating criminal or tortious conduct. In both Bills even if distributed with the requisite intention - of facilitating criminal or tortious conduct - distribution by someone who did not create it would not have attracted criminal liability, unless they had knowledge that the material was a Deepfake. On its face, the scope

of the Federal Bill was the broader of the two, criminalising those committed the *actus reus* with the intention of facilitating anything that would be a crime or tortious wrong under either Federal, State, local or Tribal law. At the time of writing, the Massachusetts Bill had been reintroduced and was under consideration.[522]

(iii) Congressional Bill 3230 – The DEEPFAKE Accountability Act

The most comprehensive set of measures in any proposals for legislation, can be found in another Bill that was introduced in the U.S. Congress in 2019 – 'Defending Each and Every Person from False Appearances and Keeping Exploitation Subject to Accountability 'DEEP FAKES Accountability' Act.[523] this also failed to pass. The preamble to the proposed legislation stated that it was intended 'to combat the spread of disinformation through restrictions on deep-fake video alteration technology'. It sought to achieve this aim in a number of ways, including amendment of title 18 of the United States Code to create a number of new criminal offences, and by establishing a number of causes of action in private law.

(a) Transparency Requirements

The legislation would have imposed transparency requirements on those who produce Deepfakes. Specifically, the provisions would have required any person who produces an 'advanced technological false personification record' (Deepfake)[524] with the intention of distributing it over the internet, or in the knowledge that it was to be distributed in this way, a duty to comply with 'watermark' and disclosure requirements. These were intended to ensure that anyone who might see a Deepfake would be aware that what they were being presented with were manipulated (or altered) images or audio. The Digital Watermark provision would have required any Deepfake which contained 'a moving visual element' to contain an embedded digital watermark clearly identifying it as containing altered visual or visual material. The legislation would also have required the Attorney General to draft rules governing the technical specifications of the digital watermarks that were required.

In addition to the watermark, the proposed legislation would have required the Deepfake to include – depending on the nature of the material and what had been altered – either an 'audio-visual-', visual-' or 'audio disclosure'. Where an audio-visual Deepfake had been created, the legislation would have required the material to contain both (i) 'a clearly articulated verbal statement that identifies the record as containing altered audio and visual elements, and a concise description of the extent of the alteration', and (ii) an unobscured written statement in clearly readable text appearing at the bottom of the visual element identifying the material as containing altered audio and visual elements, and a concise description of the extent of such alteration.' In respect of Deepfakes that contain no manipulated audio, only manipulated visual material, the requirement would have been for a written statement (a 'visual disclosure'). For that containing only audio, only a clearly articulated verbal statement would be required (an 'audio disclosure'). Anyone who failed to comply with the Watermark and disclosure requirements would have been liable for a civil penalty of up to $150,000. So too, anyone who altered a Deepfake to obscure or remove any of the required disclosures that had been included in the material

with the intention of distributing it.
There would have been a number of exceptions to the disclosure requirements: (i) where the person who created the Deepfake had included a disclosure statement which a reasonable person would consider was more prominent than that required by the proposed legislation, (ii) where the Deepfake primarily contained images and sound recordings of actual persons, e.g performing artists, (iii) images created in the editing of a motion picture, television, music or similar performance, and (iv) where the Deepfake appeared in a context such that a reasonable person would not mistake it for the actual material activity of a living person, (v) material produced by an officer or employee of the United States in furtherance of public safety or national security.

To advance the aim of these proposals, there were provisions in the Bill, the effect of which would have been to require those manufacturing software that it could reasonably be assumed would be used to make Deepfakes, to ensure that their products enabled the required watermarks and disclosures to be inserted. Further, the terms of use of these products would have had to make it clear to the purchasers, what their legal obligations were under the proposed legislation.

(b) Criminal Offences

The structure of the proposed legislation was such that failure to comply with the watermark and disclosure requirements became the *actus reus* element of a number of criminal offences. These would have been distinguished from one another by the *mens rea* elements of each. The offences would have been committed by those who failed to comply with the digital watermark and disclosure requirements:

(i) with the intention of humiliating or otherwise harassing the person falsely exhibited, provided the Deepfake record contained sexual content and appeared to feature the person engaging in sexual acts of in a state of nudity;

(ii) with the intention of causing violence or physical harm, inciting armed or diplomatic conflict, or interfering in an official proceeding, including an election, provided the Deepfake did in fact pose a credible threat of instigating or advancing such; or

(iii) in the course of criminal conduct related to fraud, including securities fraud and wire fraud, false personation, to identity theft.

An offence would also have been committed by a foreign power, or an agent of such a power, 'with the intention of influencing a domestic public policy debate, interfering in a Federal, State, local or territorial election, or engaging in other action that such a power could not lawfully undertake.

Under the legislation proposed in this Bill, criminal liability arising from the deceptive use of a Deepfake would have been generally far more extensive than that proposed by any other Bill, or any law that has so far been enacted. In respect of political interference, it would have criminalised not only those who distribute Deepfakes in the period leading up to an election – i.e. during campaigning – with the intention of harming a candidate or influencing voters, but also on those who also seek to influence wider political discourse and policy-making. However, these offences could only have been committed by foreign powers and their agents.

(c) Private Law Action

In addition to the criminal offences it would have created, the Bill would also have established for anyone depicted engaging in falsified activity in a Deepfake that did not comply with the watermark and disclosure requirements, a cause of action in private law. It would have been possible to commence such an action against the person who created the Deepfake in violation of these requirements, or who removed or obscured a disclosure that had been included in it when created. The legislation would have provided for various levels of damages: $50,000 where the person depicted in the Deepfake experienced a perceptible individual harm or there was a tangible risk of experiencing such harm; $100,000 where the Deepfake purported to depict extreme or outrageous conduct by the person, and; $150,000 where the it falsely depicted the person in sexual activity or a state of nudity, and the intention was to humiliate or otherwise harass them.

(d) Privacy Protections

As with the Californian Bill referred to in Part IV(ii) above, this Bill would have provided privacy protections for those depicted in Deepfakes that were the subject of a private law action, or proceedings commenced by Federal authorities. In respect of the latter, the legislation would have required Federal authorities to consult with people depicted in the Deepfake that gave rise to proceedings, regarding measures that the authorities might reasonably undertake to protect their privacy. In the case of a private action, provision would have been made for documents to be 'filed under seal', that is to say, filed with the court without them becoming part of the public record and generally accessible as such. Unlike the Californian Bill, the proposed legislation made clear the reasons for these privacy provisions. The provision relating to proceedings initiated by Federal authorities explained that privacy-protecting measures were intended to 'minimise additional public viewings of the Deepfake material.' That relating to private actions stated that the plaintiff could file documents under seal if there would have been a reasonable likelihood that the creation of public records regarding the Deepfake 'would result in embarrassing or otherwise harmful publicization of the falsified material activity in the Deepfake'.

(e) Institutional Research and Support

The Bill contained a number of measures that were intended to support its overall aim of 'combatting the spread of disinformation'. These included the establishment of a Deep Fakes Task Force within the Department of Homeland Security's Science and Technology Directorate. It would have been responsible for: advancing the efforts of the United States Government to combat the national security implications of Deep Fakes; researching and developing technologies to detect and counter/combat Deep Fakes and other advanced forms of image manipulation, and that would be capable of distinguishing deep fakes and similar forgeries from legitimate audio-visual recordings; providing administrative and scientific support to other Federal bodies that were researching Deepfakes; facilitating discussion and co-operation between the U.S. Government and private sector technology enterprises, academic and research institutions, regarding the identification of Deep Fakes.

In the event that had any research undertaken by any U.S. Government body led to technology that

could be used to reliably detect Deep Fakes, and/or to distinguish them from authentic recordings, the proposed legislation would have required the President to endeavour to make the technology available to U.S. private sector internet platforms, including social networks. However, there would have been no such obligation where sharing the technology was contrary to the national interests of the United States.

(f) Reporting, and Notification Requirements

Finally, the proposed legislation also contained provisions that would have imposed an obligations on the Secretary for Homeland Security to report annually on the activities of the Deep Fakes Task Force, technological progress on measures to detect and counter Deep Fakes, and the threats to national security posed by Deep Fakes, including any efforts made by Russia, China and other states and groups.[525]

In addition, the Secretary of Homeland Security would have been required to provide the House of Representatives' Committee on Homeland Security with notification of any known attempts to interfere in an official proceeding – including elections – by any foreign state.[526]

7. Conclusion

The proportion of jurisdictions in the United States that have taken steps to regulate the Deepfakes is relatively small. Most of the legislation that has been enacted to date, addresses what seem to be perceived at the present time, to be the most problematic uses of Deepfake technology – creation and dissemination of sexually explicit material without the consent of the person depicted in it, and the use of Deepfakes to influence the outcomes of elections. However, the volume of legislation that has made it onto the statute books, is matched by that of Bills that have failed to pass. The proposals set out in these Bills have generally been more ambitious, recognising that Deepfakes might present a more pervasive threat than the legislation so far enacted might suggest. It may well be that some of what is proposed will eventually become law. The law-making process in many jurisdictions requires passage of a Bill in a relatively short period, and the demands on the legislative timetable are many. Some of the legislation that has so far been enacted had been introduced in earlier Bills that died having failed to pass the all stages of the legislative processes required for proposals to become law. It remains to be seen whether the more expansive proposals that have so far failed to pass will be enacted. If the circumstances of politics in the United States, and free speech jurisprudence, present insurmountable legislative obstacles, some of these proposals might hold appeal in jurisdictions that differ in their political culture and social values.

ENACTED LEGISLATION

Statute

Texas Senate Bill No. 751

Section 1
255.004 (d) Election Code

Deep Fake: Definition

Section 1(e)"Deepfake Video" means a video, created with the intent to deceive, that appears to depict a real person performing an action that did not occur in reality.

Law/ Activity

Election Law/ creates a criminal offence

Note:This provision is one of a number of offences found in section 255.004 – all of which are intended to maintain the integrity of the election process – see other offences

Deep Fake: Definition

Section 1(e)"Deepfake Video" means a video, created with the intent to deceive, that appears to depict a real person performing an action that did not occur in reality.

Law/ Activity

Election Law/ creates a criminal offence

Note:This provision is one of a number of offences found in section 255.004 – all of which are intended to maintain the integrity of the election process – see other offences

Offences/Cause of Action

Section 1(d)
Actus Reus:
(1) Creates a deep fake video, and
(2) Causes that video to be published within 30 days of an election

Mens Rea: With intention to injure a candidate or influence the result of an election

Defences/Limitation of liability

No specific defences

Penalty/Remedy

Class A Misdemeanor
Most serious of lower level offence

Max penalty: $4000 fine and/or 1 year imprisonment

Sponsor/Committee notes

"As this technology continues to advance, it will become increasingly difficult to trust anything in public discourse"

"This technology likely cannot be constitutionally banned altogether, but it can be narrowly limited to avoid what may be its greatest potential threat: the electoral process"

Statute

California
Assembly
Bill No.730

Deep Fake: Definition

'Materially deceptive audio or visual media':

An image or an audio or video recording of a candidates appearance, speech, or conduct that has been intentionally manipulated in manner such that both of the following conditions are met:

1) The image of audio or video recording would falsely appear to a reasonable person to be authentic;
2) The image or audio or video recording would cause a reasonable person to have a fundamentally different understanding or impression of the expressive content of the image or audio or video recording than that person would have of the person were hearing or seeing the unaltered, original version of the image or audio or video recording.

Law/ Activity

Election Law/Private Law

Offences/Cause of Action

Creates statutory cause of action for available to a candidate for elective office against a person or committee who within 60 days of an election at which the candidate will appear on the ballot, distributes with actual malice, 'materially deceptive audio or visual media', with intent to injure the candidates reputation or to deceive a voter into voting against the candidate.

Defences/Limitation of liability

No liability if the audio or visual media includes a disclosure state that the image/audio./video has been manipulated.

The following are exempt from liability:
(1) Broadcast news rep[orts or documentaries provided they clearly state the media is deceptive;
(2) Broadcast media that is paid to broadcast the materially deceptive material
(3) Internet sites that regularly publishes news-content provided it is clearly stated that the material content does not accurately represent the speech or conduct of the candidate

Does not negate rights obligations of immunities under s.230 Title 47 USC

Penalty/Remedy

Candidate for office may seek injunctive or other equitable relief prohibiting distribution of the audio or visual media (s.4(3)(c)(1))

Candidate may bring action for specific or general damages against person, committee, or other entity that distributed the audio visual material.

Sponsor/Committee notes

-

Statute

California
Assembly Bill No.602
Adds new section to Civil Code
Section 1708.86
In Force: October 3, 2019

Deep Fake: Definition

The legislation creates a cause of action for 'depicted individuals'

'Depicted individual' is a person who appears as a result of 'digitization' to be giving a performance they did not actually perform, or to be performing in an 'altered depiction'o

'Digitization' means to realistically depict:
a) Nude body parts of another human being as the body parts of the depicted individual
b) Computer generated nude body parts as the nude body parts of the depicted individual
c) The depicted individual engaging in sexual conduct in which the depicted individual did not engage

'Altered depiction' means a performance that was actually performed by the depicted individual but was subsequently altered to be in violation of this section

Law/ Activity

Private Law

Offences/Cause of Action

Creates Statutory Cause of Action for depicted individuals

against people who:

(1) Create and intentionally disclose sexually explicit material who know or should reasonably have known that the depicted individual in the material, did not consent to its creation or disclosure;

(2) Intentionally discloses sexually explicit material that the person did not create and the person knows the depicted individual in that material did not consent to the creation of the sexually explicit material

'Consent' – an agreement written in plain language signed knowingly and voluntarily by the 'depicted individual' that includes a general description of the sexually explicit material and the audiovisual work in which it will be incorporated.

Consent can be rescinded within 3 business days of it being given, by means of written notice

'Sexually explicit material' – any portion of an 'audiovisual work' that shows the 'depicted individual' performing in the 'nude' (defined in [10]) or appearing to engage in, or being subjected to, subject to, 'sexual conduct' (defined in [13])

Defences/Limitation of liability

The section provides a number of specific defences (in (14)(c)(1)(A):

(1) Reporting unlawful activity;
(2) Exercising a lawful duty;
(3) Hearings, trials, or other legal proceedings;

and (14)(c)(1)(B) The material is:

(1) A matter of legitimate public concern;
(2) A work of political or newsworthy value or similar work (note material is not newsworthy only because it depicts a public figure);
(3) Commentary, criticism or disclosure that is otherwise protected by the California or US Constitutions

It is no defence to include a disclaimer in the sexually explicit material that the inclusion of the depicted individual was unauthorized, or that the depicted individual did not participate in the creation or development of the material

Statutory limitation period – 3 years from when creation, development or disclosure was discovered or should have been discovered with exercise of due diligence

Penalty/Remedy

(1) An amount equal to the monetary gain made by D from the creation, development, or disclosure of the sexually explicit material;

And either:

(a) Economic and non-economic damages caused by disclosure of the sexually explicit material, including emotional distress, or
(b) Statutory damages for any unauthorised act with respect to any one work, as follows:
- $1,500-$30,000
- unlawful act committed with malice up to $150,000-
Punitive damages - Legal fees and costs
- Any other available relief including injunction

These remedies are cumulative, and do not restrict remedies available under other law

'Malice' – D acted with intent to cause harm to P or despicable conduct done willfully with disregard for rights of P.

Sponsor/Committee notes

Note that earlier Bill that is substantially similar to this Bill (SB 584) includes provisions concerning plaintiff anonymity that have been omitted

Statute

Virginia
Section
18.2-386.1
Unlawful creation of image
18.2-386.2
Unlawful sale/dissemination of image

Deep Fake: Definition

'any videographic or still image created by any means whatsoever that depicts another person'

'Another person' – includes a person whose image was used in creating, adapting, or modifying a videographic or still image with the intent to depict an actual person and who is recognizable as an actual person by the person's face, likeness or other distinguishing feature

Law/ Activity

Criminal Law - Non-consensual creation and dissemination of sexual images

Offences/Cause of Action

18.2-386.1
Unlawful Creation(?)

Actus Reus:
Creating a videographic or still image by any means whatsoever of any nonconsenting person if (i) person is totally nude, clad in undergarments, or in a state of undress so as to expose the genitals, pubic area, buttocks or female breast in a restroom, dressing room, locker room, hotel room, motel room, tanning bed, tanning booth, bedroom or other location

Mens Rea:
Knowingly and intentionally fulfilling doing the acts that constitute the actus reus
(although its possible – not sure the above provision intended to regulate deepfakes)

18.2-386.2
Unlawful sale/dissemination of image

Actus reus:
Disseminate or sell any videographic or still image created by any means whatsoever that 'depicts another person who is totally nude or in a state of undress so as to expose genitals, pubic area buttocks, or female breasts

Mens rea:
Dissemination or sale must be with intention to coerce, harass or intimidate, and done maliciously

Knowledge that there is no authorization or licence to disseminate or sell such videographic or still image

Defences/Limitation of liability

18.2-386.1
Unlawful Creation

Recording by law enforcement and prison officials involved in investigation

18.2-386.2
Unlawful sale/dissemination of image

Para B exempts any ISP and electronic mail service provider, or any other information service, system, or access software provider that enables access to a server from multiple computers, from criminal liability where a person has used their services to disseminate the videographic or still image.

Penalty/Remedy

18.2-386.1
Unlawful Creation

Class 6 Felony

Fine $2500 and/or
Imprisonment max 12 months

18.2-386.2
Unlawful sale/dissemination of image

Class 1 Misdemeanor

Fine $2500 and/or
Imprisonment max 12 months

Sponsor/Committee notes

See notes on scope of the offence of creation – seems to be limited to voyeuristic recording – but possibly extends to fake voyeuristic material using some material recorded in prescribed circumstances

Statute

New York
S 5959-D
Amends NY Civil Rights Law, Article 5 – inserts new
Section 50-f (Right of Publicity)
and
Section 52-c (Private of Right of Action for Unlawful dissemination or publication of a sexually explicit depiction of an individual)

Deep Fake: Definition

Section 50-f (c) – Publicity Rights:

'Digital replica' – means 'a newly created, original, computer-generated, electronic performance by an individual in a separate and newly created expressive sound recording or audiovisual work in which the individual did not actually perform, that is so realistic that a reasonable observer would believe it is a performance by the individual being portrayed and no other individual.
A digital replica does not include the electronic reproduction, computer generated or other digital remastering of an expressive sound recording or audiovisual work consisting of an individual's original or recorded performance, nor the making or duplication of another recording that consists entirely of the independent fixation of other sounds, even if such sounds imitate simulate the voice of the individual.

Section 52-c – Unlawful dissemination sexually explicit material:
"depicted individual" - individual who appears, as a result of digitization, to be giving a performance they did not actually perform or to be performing in a performance that was actually performed by the depicted individual but was subsequently altered

"digitization" - to realistically depict (i) the nude body parts of another human being as the nude body parts of the depicted individual, or (ii) computer-generated nude body parts as the nude body parts of the depicted individual or (iii) the depicted individual engaging in sexual conduct, as defined in subdivision ten of section 130.00 of the penal law, in which the depicted individual did not engage.

Law/ Activity

Private law

Defences/Limitation of liability

Section 50-f (c) – Publicity Rights:
Right to publicity is a property right that can be transferred in various ways – gift, contract, will ect

Creates cause of action (i) against any person who uses deceased performer's digital replica in scripted audiovisual work as a fictional character, or for the live performance of a musical work (ii) without prior consent of person who possesses property right in ceased person's digital replica, (iii) if the use is likely to deceive the public into thinking it was authorized by the person who possesses those rights

"Deceased performer" - deceased natural person domiciled in NY at the time of death who, for gain or livelihood, was regularly engaged in acting, singing, dancing, or playing a musical instrument.

Section 52-c – Unlawful dissemination sexually explicit material:
Depicted individual has cause of action against person who, (i) discloses, disseminates or publishes sexually explicit material related to the depicted individual, and (ii) person knows or reasonably should have known the depicted individual in that material did not consent to its creation, disclosure, dissemination, or publication.
Consent- Depicted individual can only consent to the creation, disclosure, dissemination, or publication of sexually explicit material by 'knowingly and voluntarily' signing agreement written in plain language. This must include general description of the sexually explicit material and audiovisual work in which incorporated.
Agreement must be provided to person depicted 3 days prior to signing so they can review terms
Consent can be rescinded within 3 business days of signing

Penalty/Remedy

Section 50-f (c) – Publicity Rights:
Cause of action exists to 40 years after performer's death.

Various uses do not give rise to liability,- satire parody – news reports, public affairs programmes, bibliographical works, sports programmes

Section 52-c – Unlawful dissemination sexually explicit material:
No liability in respect of disclosure in reporting unlawful activity or any subsequent law enforcement activity or legal proceedings; as part of a report of legitimate public concern, political or newsworthy value

Limitation on action - later of:
3 years after dissemination or publication; or
1 year from the date person discovers, or reasonably should have discovered, dissemination or publication

Sponsor/Committee notes

Section 50-f (c) – Publicity Rights:
Cause of action exists to 40 years after performer's death.

Various uses do not give rise to liability,- satire parody – news reports, public affairs programmes, bibliographical works, sports programmes

Section 52-c – Unlawful dissemination sexually explicit material:

Injunctive relief, punitive damages, compensatory damages, court costs, attorney's fees.

BILLS THAT HAVE DIED/EXPIRED

Bill

US Senate
SB 3805
Introduced Dec 2018

Deep Fake: Definition

'Deepfake' means 'an audiovisual record created or altered in a manner that the record would appear to a reasonable observer to be an authentic record of the actual speech or conduct of an individual

Law/ Activity

Criminal law: would have added a new section 1041 – 'Fraud in connection with audiovisual records – to Ch47 18USC

Offences/Cause of Action

Actus Reus: unlawfully use any means or facility of interstate or foreign commerce to create a deepfake with

Mens rea: intention to distribute and intention that the distribution will facilitate criminal or tortious conduct under Federal, State, local or Tribal law

OR

Actus Reus:
Distribute an audiovisual record

Mens Rea: with actual knowledge that the audiovisual record is a deepfake; and

With intention that the distribution of the audiovisual record would facilitate criminal or tortious conduct under Federal, State, local or Tribal Law.

Defences/Limitation of liability

Provider of an interactive computer service not liable for any action voluntarily taken in good faith to restrict access to or availability of deep fakes or to make available information to content providers or other persons, the technical means to restrict access to deep fakes

"No person shall be held liable under this section for any activity protected by the First Amendment to the US Constitution.

Penalty/Remedy

Fine or 2 years imprisonment

....

Fine or imprisonment for max 10 years

Where creation, reproduction, distribution could

(a) reasonably be expected to affect the conduct of any judicial proceeding of a Federal, State, local or Tribal government agency, including the administration of an election or the conduct of foreign relations; or

(b) facilitate violence

Notes

-

Bill

US House of Representatives

HR 3230

Too extensive to include here – see report

Bill

California
AB 1280 – Crimes Deceptive Recordings
Introduced Feb 21, 2019

Deep Fake: Definition

'Deepfake' means 'any audio or visual media in an electronic format, including any motion picture film, video recording, or sound recording that is created or altered in a manner that it would falsely appear to a reasonable observer to be an authentic record of the actual speech or conduct of the individual depicted in the recording.'

Law/ Activity

Criminal Law – would have added a new section (section 644) to the Penal Code

Offences/Cause of Action

Would have created 3 criminal offences:

1) Actus Reus:, without the individual's consent, prepares, produces or develops any deepfake that depicts an individual personally engaging in sexual conduct AND distributes or exhibits it to, exchanges it with others

Mens Rea: 'knowingly' performing the actus reus

2) Actus Reus: (i) Prepares, produces or develops any deepfake (ii) that depicts an individual under 18 years of age (ii) personally engaging in sexual conduct AND (iv) distributes or exhibits it to, or exchanges it with others, or offers to do any of these things.

Mens rea: Doing so knowingly

3) Actus reus (i) Within 60 days before an election, (ii) without consent of the depicted individual, (iii) prepares, produces or develops any deepfake AND (iv) distributes/exhibits/exchanges deepfake to/with others

Mens rea: Intention that the deepfake coerce or deceive any voter into voting for or against a candidate or measure in that election

Defences/Limitation of liability

Offence 1:
$1000 fine or 1 year imprisonment or both

Offence 2:
$10,000 fine or 1 year imprisonment, or both

Offence 3:
$1000 fine or 1 year imprisonment or both

Penalty/Remedy

No one shall be liable for activity protected by First Amendment to US Constitution

Notes

The legislation would also have secured $25m for the University of California, from Government Funds, for research to identify and combat the inappropriate use of deepfake technology

Bill

Massachusetts
House Bill No.3366
Introduced:
17 Jan 2019

Deep Fake: Definition

'Deepfake' – 'an audiovisual record created or altered in a manner that the record would falsely appear to a reasonable observer to be an authentic record of the actual speech or conduct of an individual'

'Audiovisual record' – any audio or visual media in an electronic format, including any photograph, motion-picture film, video recording, electronic image, or sound recording.'

Law/ Activity

Criminal Law

Offences/Cause of Action

Would have created 2 offences

Offence 1:
Actus reus: Creating a deepfake

Mens Rea: (i) with the intention to distribute, and (ii) with the intent that the distribution of the deepfake would facilitate criminal or tortious conduct

Offence 2:
Actus Reus: Distributing an audio visual record

Mens Rea: (i) with actual knowledge that the audiovisual is a deepfake, and (ii) with the intention that the distribution of the audio visual recording will facilitate criminal or tortious conduct.

Defences/Limitation of liability

Fine $5,000, or 2.5 years imprisonment, or both

Penalty/Remedy

No liability for any activity protected by the Massachusetts Constitution or the First Amendment to the US Constitution.

Notes

-

Bill

New York
AB 8155
Introduced May 31, 2018

Deep Fake: Definition

'Digital replica' – means 'a computer-generated or electronic reproduction of a living or deceased individual's likeness or voice of the individual being portrayed. A digital replica is included with an individual's portrait.

Law/ Activity

Private Law/ Image (Publicity/Privacy) Rights

Offences/Cause of Action

Provides that a person has privacy and publicity rights in respect of their persona.

'Persona' – means name, portrait or picture, voice or signature of an individual.

Provides a cause of action for unauthorized use of a person's persona.

Publicity rights are a form of personal property, that can be transferred and assigned in the person's will. They continue to exist for 40 years after the person's death, and can be enforced by any person in who they vest, e.g. relatives.

There are particular provisions concerning commercial use of digital replicas in dramatic, musical and sporting performances. These are violations of publicity rights if used without the consent of the individual or other publicity rights-holder

Defences/Limitation of liability

Injunctions to prevent unauthorized use.
Damages for unauthorized use.

Penalty/Remedy

Various exceptions include use of digital replica for purposes of:

parody, satire, commentary and criticism;

political; works, works of public interest, newsworthy value, documentary

de minimis or incidental use

Notes

-

BILLS THAT HAVE DIED/EXPIRED

Bill

HR 1
Introduced:
Jan 4, 2021

Deep Fake: Definition

The term 'materially deceptive audio or visual media' means 'an image or an audio or video recording of a candidate's appearance, speech or conduct that has been intentionally manipulated in a manner such that both of the following conditions are met:
(1) The image or audio or video recording would falsely appear to a reasonable person to be authentic
(2) The image or audio or video recording would cause a reasonable person to have a fundamentally different understanding or impression of the expressive content of the image or audio or video recording than that person would have of the person were hearing or seeing the unaltered, original version of the image or video recording.

Area of Law/Activity

Electoral Law/Private Law/Criminal Law

Would amend Title III Federal Election Campaign Act 1971 (52 USC 30101 et seq) to introduce new section 325

Offences/Cause of Action

General Prohibition:
Person, political committee or entity, within 60 days of election for Federal office, must not distribute with actual malice, materially deceptive audio or visual media of a candidate with intention of injuring the candidate's reputation or to deceive a voter into voting for or against the candidate.

Civil Action:
Provides cause of action and remedies for a candidate for office whose voice or likeness appears in a materially deceptive audio or visual media distributed in violation of general prohibition

Crime:
Wilfully and knowingly violates the general prohibition.
In an action for defamation - Violation of the general prohibition constitutes defamation per se.

Defences/Limitation of liability

No prohibition if the audio or visual media includes a disclosure state 'This image/video/audio has been manipulated'

Legislation would prescribe the form the disclaimer would have to take – in terms of visibility and prominence.

Provision inapplicable to various entities:

a) Radio television broadcasting operator, programmer, producer broadcasting the media as part of a news report or documentary – providing it discloses that there are qns about the authenticity of the deceptive audio or visual media

b) Radio or television broadcaster etc, that is paid top broadcast the materially deceptive audio or visual media;

c) Internet website rehgularly published newspaper, magazine or other periodical , including internet publications that routinely carry news and commentary of general interest - providing it discloses that there are qns about the authenticity of the deceptive audio or visual media

Provision would not affect the rights, obligations and immunities of interactive service providers under s.230 Title 47 USC

Penalty/Remedy

Civil Action:

Candidate may seek (i) injunctive or other equitable relief, and (ii) bring an action in general or special damages against the person, committee, or entity that distributed the material.

Criminal Penalty:
Fine up to $100,000 Imprisonment up to 5 years, or both

Notes

-

Mass.

Bill H.1755
Introduced 29 March 2021

See Mass. H3366
- Reintroduction of substantially same bill

10.2 Interviews

10.2.1 Altena Davidsen

Interview met: **Judit Altena Davidsen**
Interview door: Bart van der Sloot
& Yvette Wagensveld
Datum: 14 06 2021

Het Nederlandse bewijsstelsel kan worden gekarakteriseerd als 'negatief-wettelijk', hetgeen betekent dat de rechter bij een gebrek aan voldoende wettig bewijs gedwongen is tot een vrijspraak, maar bij aanwezigheid van voldoende wettig bewijs niet gedwongen is om tegen zijn overtuiging in tot een veroordeling te komen. De bewijsstandaard is dat het tenlastegelegde feit wettig en overtuigend moet zijn bewezen. De heersende opvatting is dat dit betekent dat de schuld van de verdachte buiten redelijke twijfel moet vaststaan op grond van door de wet erkende bewijsmiddelen.

De wet somt op van welke bewijsmiddelen de rechter gebruik kan maken. Deze opsomming gaat over de vorm waarin het bewijs is gegoten. Videobeelden vallen daar tot op heden niet onder, maar worden via een beschrijving van de inhoud in een ambtsedig opgemaakt proces-verbaal of via de eigen waarneming van de rechter gebruikt voor het bewijs. Ondanks de binding aan de wettige bewijsmiddelen heeft de strafrechter een grote vrijheid in de selectie en waardering van het bewijs, die hij in het beginsel ook niet hoeft te motiveren. In algemene zin kun je wel zeggen dat bewijsstukken rechtmatig moeten zijn vergaard, betrouwbaar moeten zijn en bovendien relevant voor het bewijs van het tenlastegelegde. De wet en jurisprudentie sluiten op basis van deze eisen enkele bewijsmiddelen uit voor het bewijs. Tot slot formuleert de wet enkele bewijsminima om de deugdelijkheid van de bewijsbeslissing te bevorderen. De belangrijkste daarvan is de regel dat een bewezenverklaring niet mag worden gebaseerd op één getuigenverklaring.

Er is geen algemeen toetsingskader voor de beoordeling van de betrouwbaarheid van bewijsmiddelen: de rechter kan alles voor het bewijs gebruiken dat hem 'uit het oogpunt van betrouwbaarheid dienstig voorkomt', aldus de Hoge Raad. Enkele bewijsmiddelen zijn niettemin met het oog op de betrouwbaarheid uitgesloten van het bewijs door de Hoge Raad. Dit geldt bijvoorbeeld voor verklaringen verkregen op basis van hypnose of met een leugendetector en voor de mondelinge slachtofferverklaring op zitting en de vordering van de benadeelde partij.

De strafrechter redeneert altijd met onzekerheden: op basis van uiteenlopende waarnemingen doet hij een uitspraak over het tenlastegelegde, maar honderd procent zekerheid heeft hij nooit. Een belangrijke moeilijkheid is dat niet altijd duidelijk is of en in hoeverre een bewijsmiddel betrouwbaar is: een getuige kan liegen, maar ook een verkeerde waarneming hebben gedaan of zich het gebeurde verkeerd herinneren. Ook van technisch bewijs kan de betrouwbaarheid worden betwist: video- en geluidsopnames kunnen bijvoorbeeld worden gemanipuleerd. De rechter zal dan al het beschikbare bewijsmateriaal in samenhang moeten beoordelen om tot een oordeel te komen. Er bestaat geen algemeen kader dat bepaalt hoe de rechter van die individuele bewijsmiddelen tot een bewezenverklaring moet komen.

De rechter heeft dus een grote vrijheid in de selectie en waardering van het bewijsmateriaal, en hoeft in beginsel ook niet in het vonnis

verantwoording af te leggen van de keuzes die hij daarbij maakt (er is geen motiveringsplicht). Dat ligt anders op het moment dat de betrouwbaarheid van een bewijsmiddel door of namens de verdachte in twijfel wordt getrokken. Op het moment dat de rechter het betwiste bewijsmiddel toch gebruikt, zal hij moeten uitleggen waarom hij dat betrouwbaar acht. Het kan lastig zijn voor de verdachte om goed te onderbouwen waarom er problemen zijn met de betrouwbaarheid van een bewijsmiddel, met name wanneer het gaat om deskundigenbewijs. De verdachte kan dan verzoeken om het verrichten van een contra-expertise.

Met een verbetering van de techniek bij het maken van deepfakes en toenemend gebruik daarvan kan vaker twijfel rijzen over de authenticiteit van videobeelden. Wezenlijk nieuw is die twijfel niet: ook van andere bewijsmiddelen kan als gezegd onduidelijk zijn of en in hoeverre ze een waarachtige weergave vormen van de werkelijkheid, en die onzekerheid maakt het niet onmogelijk om dat bewijsmiddel te gebruiken voor het bewijs (al wordt de bewijswaarde daarvan wel lager). Wat wel nieuw is, is dat een veelgebruikt bewijsmiddel nu wellicht vaker ter discussie zal staan dan voorheen. In zoverre verschilt de opkomst van deepfakes enigszins van de opkomst van bijvoorbeeld nieuwe vormen van digitaal bewijs. Als de voorspellingen over de kwaliteit van deepfakes en de mate waarin die zullen worden gemaakt uitkomen, valt te verwachten dat in de rechtszaal veelvuldig discussie zal ontstaan over videobeelden en dat in die discussie mogelijk deskundigen zullen worden betrokken. De rechtspraak zal een manier moeten vinden om hiermee om te gaan.

10.2.2 Ashmann

Interview met: **Margreet Ahsmann**
Interview door: Bart van der Sloot
& Yvette Wagensveld
Datum: 14 06 2021

In het civiele recht geldt dat stellen en bewijzen in elkaars verlengde liggen, met dien verstande dat wil een partij toekomen aan bewijslevering van betwiste feiten zij voldoende moet hebben gesteld. De vuistregel is echter niet zozeer wie stelt, moet bewijzen, maar preciezer: wie een rechtsgevolg inroept, moet daarvoor (voldoende) relevante feiten stellen en bij betwisting bewijzen. Een rechtsgevolg kan ook door de gedaagde worden ingeroepen en ook dan geldt de hoofdregel van bewijslastverdeling: 'De partij die zich beroept op rechtsgevolgen van door haar gestelde feiten of rechten, draagt de bewijslast van die feiten of rechten (...)'.[527] Die regel is voor het bewijs van cruciaal belang omdat het altijd gaat om de vraag wie het *risico* moet dragen van het niet kunnen leveren van bewijs. Als de partij die de bewijslast heeft daarin niet slaagt, bijvoorbeeld bij gebrek aan of bij onvoldoende bewijsmateriaal, is het gevolg dat haar vordering wordt afgewezen. De rechter beslist dan ten nadele van de partij die de bewijslast en dus het risico heeft.

De rechter kan echter behoefte hebben aan ingrijpen in de verdeling van het bewijsrisico om een partij rechtsbescherming te bieden. Van de genoemde hoofdregel van bewijslastverdeling kan hij afwijken, wanneer 'uit enige bijzondere regel of uit de eisen van redelijkheid en billijkheid een andere verdeling van de bewijslast voortvloeit', zoals in de bijzin van artikel 150 Rv staat.

- De meest verregaande ingreep is de volledige omkering van de bewijslast en het bewijsrisico op basis van de eisen van redelijkheid en billijkheid, zij het dat deze uitzondering in de praktijk zelden wordt toegepast. De Hoge Raad wil er eigenlijk niet aan. Een voorbeeld waarin zulks wel werd aangenomen, is een geval waarin een echtpaar staande huwelijk (op voorstel van de man die kandidaat-notaris was) alsnog huwelijkse voorwaarden had opgemaakt, waarbij de man had gezegd dat er voor haar niets ten nadele zou wijzigen. Toen de man vervolgens wilde scheiden en de huwelijkse voorwaarden toch ongunstig voor de vrouw bleken te zijn, vorderde zij vernietiging van de voorwaarden op grond van dwaling. Op grond van de hoofdregel van bewijslastverdeling zou zij de feiten die daaraan ten grondslag liggen moeten bewijzen. Als na bewijslevering onduidelijkheid zou blijven bestaan wie gelijk heeft, zou haar vordering worden afgewezen omdat zij volgens de hoofdregel het bewijsrisico draagt. In dit specifieke geval werd daarom de bewijslast en het bewijsrisico omgekeerd, zodat de man het bewijs moest leveren dat zij niet had gedwaald. Zou hij niet in het bewijs slagen dan zou de vordering van de vrouw worden toegewezen, een verregaande consequentie dus. [528]
- Een andere, frequenter toegepaste maatregel die de rechter zou kunnen treffen voor een partij in bewijsnood, is een verzwaarde stelplicht, juister geformuleerd: een verzwaarde informatieplicht leggen op de partij die niet de bewijslast heeft. Deze correctie op de bewijslastverdeling past de rechter vooral toe in situaties waarbij de partij die de bewijslast heeft weinig inzicht in de feiten heeft, in tegenstelling tot de wederpartij; de rechter doet dan aan ongelijkheidscompensatie. Denk bijvoorbeeld aan een patiënt die een ziekenhuis aansprakelijk wil stellen op grond van een fout gemaakt tijdens de operatie. De patiënt roept het rechtsgevolg in - hij wil schadevergoeding - en moet daartoe dus voldoende feiten stellen. Dat is cruciaal want als hij dat onvoldoende doet, komt de rechter zelfs niet toe aan een bewijsopdracht en wordt zijn vordering direct afgewezen. Maar hoe komt de patiënt aan de feiten die zijn klacht onderbouwen, nu deze zich immers hebben afgespeeld in het domein van het ziekenhuis? Het ziekenhuis moet dan volgens vaste jurisprudentie concrete en nauwkeurige informatie verschaffen (bijvoorbeeld het patiëntendossier overleggen) over de gang van zaken tijdens de medische behandeling.[529]

Voor deepfakes laat het stelsel van bewijslastverdeling op zichzelf voldoende ruimte aan de rechter om het slachtoffer die beweert dat iets deepfake is en daardoor in bewijsnood komt, tegemoet te komen. De rechter kan ingrijpen door het bewijsrisico om te keren en het te leggen bij de wederpartij die beweert dat de videobeelden wél integer zijn, dus geen deepfake is. Die partij moet dan dus het tegendeel bewijzen, namelijk dat het beeldmateriaal echt is. Die omkering gebeurt dan op basis van de redelijkheid en billijkheid.

Waarschijnlijker is dat de rechter een minder verregaande maatregel gelast door te verlangen dat de wederpartij van het slachtoffer (controleerbare) informatie verschaft met betrekking tot de herkomst van het gemanipuleerde materiaal. Gebeurt dat onvoldoende dan kan de rechter de vordering van eiser reeds op die grond toewijzen of het standpunt van eiser voorshands bewezen achten.

De rechter kan ook, wanneer specialistische kennis nodig is voor het vaststellen van bepaalde

feiten, naast of in plaats van een getuigenverhoor een deskundigenbericht gelasten; daartoe heeft hij een grote beoordelingsvrijheid (discretionaire bevoegdheid). Om te kunnen vaststellen of het bewijsmateriaal al dan niet is gemanipuleerd, zal dat waarschijnlijk snel noodzakelijk zijn.

Wat de waardering van het in het geding gebrachte bewijsmateriaal betreft, heeft de rechter ruime mogelijkheden om achter de juistheid van de feiten te komen. In het civiele recht geldt de zogeheten vrije bewijsleer, dat wil zeggen: vrijheid voor partijen - een partij kan alles als bewijs aandragen - én vrijheid voor de rechter, te weten hij bepaalt zelf of hij het bewijs geleverd acht. Er is voor de rechter dus niet een in de wet vastgelegde maatstaf, zoals in het strafrecht. Wel moet de rechter een redelijke mate van zekerheid hebben dat de te bewijzen feiten juist zijn. Om tot die overtuiging te komen, beoordeelt hij het bewijs niet in isolement, maar bekijkt het materiaal in zijn totaliteit en betrekt daarbij alle omstandigheden van het geval.

In het geval van deepfakes is het dus aan de rechter om een oordeel te vellen over de vraag of met het beeldmateriaal is geknoeid, dus over de authenticiteit (en relevantie) van het aangedragen materiaal. Vragen over authenticiteit zijn overigens niet nieuw en hebben altijd gespeeld; denk aan een valselijk opgemaakte akte. In zekere zin is de vraag hoe met deepfake-video's om te gaan dus niet wezenlijk anders. Het gaat ook hier om het verifiëren van de totstandkoming en herkomst van (gemanipuleerd) bewijsmateriaal. De rechter zal onderzoeken of er aanleiding is te twijfelen aan de authenticiteit van de vermeende deepfake en in zijn beoordeling zo nodig ook betrekken of er andere feiten zijn waaruit de (on) waarachtigheid van de vermeende deepfake zou kunnen blijken. Certificering van voor bewijs bestemde digitale bestanden zou daaraan op voorhand tegemoet kunnen komen.

Om tot een betrouwbare bewijsbeslissing te komen, is het wel van belang dat rechters enige kennis vergaren over digitale videomanipulatietechnieken, zodat ze de hierover ingenomen stellingen en de uitgebrachte deskundigenberichten beter kunnen begrijpen.

10.2.3 Bock

Interview met: **Ruth de Bock**
Interview door: Bart van der Sloot
& Yvette Wagensveld
Datum: 21 06 2021

Binnen het civiele bewijsrecht geldt als uitgangspunt dat de rechter feiten als vaststaand aanneemt, wanneer deze door de ene partij zijn gesteld en door de andere partij niet gemotiveerd zijn betwist. De consequentie daarvan is dat de rechter in beginsel uitgaat van de authenticiteit van de schriftelijke bewijsstukken die zijn aangeleverd, zolang de wederpartij niet gemotiveerd stelt dat die stukken *niet* authentiek zijn. De rechter zal dus niet standaard en uit eigen beweging al het schriftelijke bewijs dat is aangeleverd verifiëren en beoordelen op authenticiteit. Alleen als er heel duidelijke aanwijzingen zijn dat bewijsstukken niet authentiek zijn, zal de rechter uit eigen beweging hiernaar een nader onderzoek instellen. Maar als daarvan geen sprake is zal de rechter pas 'aanslaan' als de wederpartij aanvoert dat sprake is van niet-authentieke (vervalste) stukken. Het is in die zin niet heel ingewikkeld om de rechter een rad voor de ogen te draaien: als de wederpartij niet 'aanslaat' is het maar zeer de vraag of

rechter uit zichzelf toetst of het bewijsmateriaal authentiek is.

Binnen het civielrecht geldt de vrije bewijsleer. Dat betekent dat alle bewijsmiddelen zijn toegestaan. Bovendien is er geen wettelijk kader waaraan bewijsmiddelen kunnen of moeten worden getoetst. De rechter is vrij bij de bewijswaardering; er is geen wettelijk bewijsminimum. Ook is het aan de rechter om te beoordelen of bewijsmateriaal rechtmatig, authentiek en relevant is. . Maar, zoals gezegd, aan een beoordeling van de authenticiteit van een bewijsstuk komt de rechter in het algemeen pas toe als er op dat punt gemotiveerd verweer wordt gevoerd door de wederpartij. Op zichzelf is dat ook niet zo gek, want de wederpartij is ook degene die kan beoordelen of, bijvoorbeeld, een email in het geding wordt gebracht die hijzelf nooit heeft ontvangen, waardoor er gerede twijfel is aan de authenticiteit van die email. De rechter weet dat niet.

Ook is belangrijk om te vermelden dat er voor procespartijen een waarheidsplicht geldt. 'Partijen zijn verplicht de voor de beslissing van belang zijnde feiten volledig en naar waarheid aan te voeren. Wordt deze verplichting niet nageleefd, dan kan de rechter daaruit de gevolgtrekking maken die hij geraden acht.'[530] Dit betekent dat het een partij niet is toegestaan om feitelijke onjuistheden naar voren te brengen in de procedure, Uiteraard is het daarmee ook niet toegestaan om niet-authentieke (vervalste) bewijsstukken over te leggen in de procedure.

Advocaten spelen hier in de praktijk een belangrijke rol in. Zij hebben een professionele zorgplicht om ervoor zorg te dragen dat het bewijs dat aan de rechter wordt aangedragen authentiek en deugdelijk is. Zij mogen – uiteraard – niet meewerken aan het vervalsen van bewijsmateriaal dat in de procedure wordt gepresenteerd.

Als bewijsmateriaal dat is aangedragen niet authentiek blijkt te zijn – als een van de partijen de rechter om de tuin probeert te leiden – dan zal dit nadelige gevolgen hebben voor die partij. Het kan tot gevolg hebben dat de rechtszaak in het nadeel van die partij wordt afgesloten of dat een partij in de proceskosten wordt veroordeeld. Maar het kan ook voorkomen dat er simpelweg een streep door de vordering gaat.

Ook komt het sporadisch voor dat advocaten die willens en wetens ondeugdelijk bewijs aan de rechter hebben aangeleverd tuchtrechtelijk worden aangesproken.

Ten aanzien van deepfakes zou er kunnen worden gewerkt met een authenticatiesysteem. Je mag alleen bewijs aandragen dat een authenticatiestempel heeft. Mails kunnen bijvoorbeeld vrij eenvoudig worden vervalst. Daar zou je kunnen werken met een bepaalde regel dat je mails alleen als bewijsstuk zou kunnen aandragen als het via een bepaald veilig en betrouwbaar systeem is verstuurd en opgeslagen. Zo'n soort systeem is in de toekomst denkbaar bij ander bewijsmateriaal, zoals video en audiofragmenten.

Een andere optie zou kunnen zijn een grotere zorgplicht neerleggen bij advocaten, die als poortwachter gaan functioneren. Binnen het strafrecht heb je dan de politie/het OM dat ambtenaren tot zijn beschikking heeft om verificatie van bewijsmateriaal op zich te nemen. Binnen het civielrecht zou de meest aangewezen

partij de advocaat zijn, maar het is maar de vraag of de advocaat kan worden aangesproken als cliënten de advocaat zelf om de tuin leiden en de advocaat dus niet weet dat sprake is van een vervalsing.

10.2.4 Burkell & Gosse

Interview with: **Jacquelyn Burkell & Chandell Gosse**
Interview by: Bart van der Sloot
Date: 16 06 2021

Initially, the term deepfake was used to refer to limited use cases, in particular pasting the face of one person on the body of another. Gradually, the term came to stand more broadly for synthetic media and refers now to both audio and video manipulations. One of the problems with such a broad definition is that it covers everything and is thus rendered meaningless. On the other hand, many of the problems triggered by deepfakes fit in a general trend and are similar to those triggered by manipulated images and misinformation in general as well as image abuse.

An example of where deepfakes fit in the general picture is that deepfakes are primarily, more than 90%, used for sexual/pornographic output, directed at women. As such, it echoes the societal problem, both offline and in particular online, of sexualising women's bodies and promoting an unrealistic beauty ideal for women. It also removes agency over one's own bodily representation and invades a person's sexual privacy. Another example is the use of deepfakes for misinformation. In a certain sense, deepfakes are just another tool used for creating and distributing 'fake news'.

But deepfake technology also has certain characteristics that make it stand out. It can be used for a full spectrum, multichannel representation of reality. Because of the use of machine learning technology, the manipulated audio and video content produced is increasingly difficult to distinguish from real/authentic content. The existence of these 'near real' but false productions do undermine our shared sense of reality, and the better they get, the greater that effect.

The main use cases of deepfake technology are and will be primarily negative. Porn/sexual deepfakes will remain the primary use case. Importantly, besides the well-known harms associated with sexual deepfakes, deepfakes can do harm even if the person depicted and everyone seeing the deepfake knows it to be false. For example, a sexual deepfake of a schoolgirl might have important social consequences and change interpersonal dynamics in the classroom, even if the class mates very well know that the schoolgirl did not perform the actions she appears to be performing. In addition, seeing yourself perform actions in a deepfake you know you did not perform may nevertheless change your perception of yourself. One does not need to refer to Baudrillard to understand how simulacra may come to replace/be reality; e.g. if you smile and laugh a lot, that very action may make you happy. Finally, it may not matter to people whether something is fake or not; if news fits a person's political beliefs, he/she might readily accept it, without wanting to verify its veracity. A person making a deepfake porno of a celebrity very well knows it to be fake, but does not care.

Obviously, there are potential positive use cases of deepfakes as well,[531] but the question here is what is positive? Having a politician address minority populations in their language may seem like a good idea, but it also undermines a shared sense of reality – and it isn't real. The politician may seem to know the language, but they do not. Having a historical figure teach in schools about their historical time may appeal to youngsters, but it might also just be creepy.

One of the biggest problems right now is the fact that by far most people are unaware of deepfake technology and its capacity. Even if they have heard about it, people do not generally assess content they see or verify its authenticity and trustworthiness. That is why public awareness campaigns and education in this field might be helpful. Such campaigns should steer audiences away from blaming the victims of false representations, as is often the case. Obviously, women can do a lot to protect themselves and take matters into their own hands, but this should not mean that they are responsible if something goes wrong. The main message should be that it is wrong to create non-consensual deepfakes and distribute it.

Detection technology may be able to filter out some of the deepfakes, but it will not always be successful. There will always be an arms race between technology to create deepfakes and technology to detect it, the latter will never win.

A stricter stance on some of the apps on the consumer market, such as deepnude apps[532] might help. In particular, emphasis should be put on sexual uses of deepfake technology, and not, such as is currently the case, on real or hypothetical uses for political purposes, such as having the prime minister/president/political leader give a speech he/she did not give. Not only should primary attention be given to sexual deepfakes because this represents by far most uses of this technology, but also because political leaders have sufficient resources to address potential fake content about them. This is different from the case of sexual deepfakes are created about ordinary women (i.e. non-celebrities), who generally have little understanding of deepfake technology and the complexities of the online environment and have limited recourses to start long and complicated legal proceedings.[533]

10.2.5 Dunnen

Interview door: Bart van der Sloot
Interview met: **Manon den Dunnen**
Strategisch Specialist Digitaal, Nationale Politie
datum: 08 06 2021

Onze benadering van het fenomeen Deepfake is in de loop der tijd veranderd. Aanvankelijk hanteerden we de begin vorig jaar nog gebruikelijke afgebakende definitie, gerelateerd aan het met kwade intenties, met behulp van bepaalde AI technieken, gemanipuleerd of gegenereerd beeldmateriaal. Op ieder van die punten bleek deze benadering echter te beperkt. Zeker als je in kaart wilt brengen wat de consequenties zijn voor het politiewerk. Dan is het veel relevanter om te kijken naar de grotere ontwikkeling van synthetische media.
Bij synthetische media gaat het om met (AI) gegenereerde of gemanipuleerde media. Dit kan gaan om bijvoorbeeld (bewegende) beelden zoals gezichten, geluid, waaronder stemmen, eigenlijk alle digitale informatie en signalen. Deepfakes zijn hier onderdeel van, maar omdat

de term nu zo'n hype is zie je dat mensen de term Deepfake echt overal voor toepassen, ook voor met andere vormen van AI gegenereerde *niet-bestaande* personen, in bijvoorbeeld games. En ook voor positieve applicaties, zoals het klonen van de stem van iemand met ALS of het communiceren met een Deepfake van een dierbare die is overleden bij rouwverwerking.

Volgens Nina Schick, auteur van *Deepfakes and the Infocalypse – What You Urgently Need to Know.* verwachten experts dat binnen 6 jaar meer dan 90% van de online content synthetisch is. Als je er eenmaal oog voor hebt zie je dit ook gebeuren, zichtbaar, zoals op websites voor brillen die je op een synthetische versie van je gezicht kunt passen of synthetische modellen op een lingerie site, waarvan je de huidskleur kunt aanpassen om te kijken of de kleur bij jou past. Maar het gebeurt ook ongemerkt, achter de schermen. Denk bijvoorbeeld aan software voor videobellen, waarbij de ogen van de deelnemers synthetisch zo gemanipuleerd worden dat ze standaard in de lens kijken. Of de camera in je telefoon; het begon met automatische correctie van rode ogen, maar nu wordt veel meer synthetisch gegenereerd. Dat kwam vorig jaar naar buiten toen tijdens de grote bosbranden in California mensen foto's en filmpjes wilden maken van de oranje lucht. Door de in veel telefoons ingebouwde (algoritmen ("intelligentie") werd de kleur werd automatisch aangepast van oranje naar grijs.[534]

Synthetische luchten, synthetische videocalls, allemaal "nep" dus, maar puur vanuit een verbeterde dienstverlening, nooit bedoeld voor bewijsvoering. Maar het beïnvloedt wel onze waarneming en voor de politie veel van haar werkprocessen. De trend is richting steeds meer communicatie online. Maar als we spreken met mensen kijken we ook naar non-verbale communicatie. Hoe weet je wat daar straks nog echt van is? Dat wordt afhankelijk van de gebruikte tools voor videoconferencing, van hoeveel meetpunten uit het gezicht over de lijn gaan om vervolgens aan de ontvangende kant het gezicht weer op te bouwen. Moeten we straks een kennisbank opbouwen van de aanpassingen die ingebakken zitten in de verschillende hard- en software, dat lijkt ondoenlijk?

"In de context van politiewerk bieden deepfakes/ synthetische media zowel kansen als risico's. Momenteel is de organisatie bezig met interne bewustwordingssessies als het gaat om Synthetische media en Deepfakes. Tijdens deze workshops bespreken we wat dit kan betekenen voor onze werkwijze en hoe we hierop kunnen anticiperen. Daarnaast brainstormen we over de kansen en de risico's".

Ten aanzien van mogelijk toekomstige positieve toepassingen voor de politie valt te denken aan:

- 1. Chatbots, virtuele agenten
- 2. Datasets genereren voor het testen van AI (bijvoorbeeld herkennen objecten in foto's) en diversiteit in bijv. gebruikte foto's
- 3. Toegankelijkere (opsporings) communicatie/voorlichting, in eigen taal[535] en/of door aansprekende personen;[536]
- 4. Gebruik van Deepfake gezichten om de authenticiteit van fake profielen te vergroten, bijvoorbeeld zoals recent toegepast is in een mensensmokkelzaak.[537] In de toekomst kun je hier ook synthetische stemmen en tekstsynthese bij gebruiken.
- 5. Mogelijkheden voor het afschermen van getuigen[538] en eigen medewerkers[539]
- 6. Etc.

Ditsoortideeënstaannunogindekinderschoenen. Er is een aantal obstakels voordat met dit soort zaken kan worden geëxperimenteerd, waarvan naast organisatorische en technische drempels, de belangrijkste in het juridische en ethische domein gelegen zijn. Juridisch is veel nog niet uitgekristalliseerd in wetgeving en jurisprudentie. Daarom wordt nu bijvoorbeeld eerst afgewacht om te weten hoe de rechter tegen de toepassing in de mensensmokkelzaak aankijkt. Daarnaast is in september een sessie gepland met hoogleraren, rechters en anderen om te reflecteren op waar het Wetboek van Strafvordering en het Wetboek van Strafrecht mogelijk nog niet in voorziet.

Ten aanzien van de ethische kant speelt een aantal meer fundamentele vragen, zoals in hoeverre moet je als politie bijvoorbeeld bijdragen aan de groeiende hoeveelheid 'nepcontent' en de vermenging daarvan met de echte wereld. Ondermijnt dit wellicht het vertrouwen in de digitale samenleving/economie en in de politie?

Ten aanzien van de gevaren van Deepfakes binnen het politiedomein kan een onderscheid worden gemaakt in drie typen risico's:

- 1. Groei criminaliteit

"We kunnen een exponentiele stijging verwachten van criminaliteit gepleegd door middel of met behulp van Deepfakes c.q. synthetische media; denk daarbij aan (identiteits)fraude, sextortion, etc. Europol[540] heeft hier een goed rapport over geschreven, zij benadrukken ook hoe met tekstsynthese de eerste stappen in social engineering, phishing etc. volledig geautomatiseerd kunnen worden".

- 2. Informatiepositie

"We zijn een informatiegestuurde organisatie waarbij we ons handelen voor een groot deel baseren op informatie van buiten, van organisaties, burgers maar ook van het internet. Hoe goed hebben die organisaties hun informatieproces onder controle. Denk aan het simpele voorbeeld van identiteitspapieren, mensen mogen zelf een pasfoto aanleveren. Als je weet dat het al mogelijk is om een foto zodanig te bewerken is dat die voor mensen op persoon A lijkt en voor de machine op persoon B (gezichtsherkenning), dan is dat een risico.
Daarnaast hebben nepberichten impact als sturingsinformatie, denk aan het voorbeeld van het nepbericht waarin het leek dat RTL Nieuws een aankondiging deed van een bommelding rondom een plein in Heerlen. In dat geval kon het bericht snel weerlegd worden, maar vanwege de ontstane paniek moet je toch handelen."

- 3. Waarheidsvinding

" Denk aan fake alibi's, maar ook de authenticiteit van het materiaal dat de politie binnenkrijgt over mogelijke delicten. Als veruit de meeste content synthetisch van aard is, dan is bijna alles deels of geheel gemanipuleerd. Je kunt verwachten dat in de rechtszaal door verdachten of hun advocaten steeds vaker geroepen zal worden dat iets 'fake' is of gemanipuleerd. De politie zal dus op een transparante manier moeten onderbouwen waarom de betreffende digitale informatie welke bijdrage heeft in de bewijslast en hoe de 'echtheid' forensisch is geborgd. Dus de mate waarin de manipulatie wel/niet relevant is in de context van het onderzoek. Hier is al ervaring mee, regelmatig moeten digitaal rechercheurs onderbouwen dat bewijsmateriaal niet door hackers op de computer van een verdachte is gekomen."

Maar de vraag is of we dit straks in alle zaken moeten doen, dit zal veel extra capaciteit kosten, maar je wilt niet dat het strafrecht ten onrechte tegen iemand wordt toegepast.

"Ik verwacht ook dat we in de toekomst met waarschijnlijkheidspercentages zullen gaan werken. Hierbij zullen we vaker een beroep op het NFI moeten doen. Met betrekking tot DNA bewijsmateriaal wordt nu al met twee percentages gewerkt, maar uit de praktijk blijkt dat rechters niet altijd in staat zijn die percentages goed op waarde te schatten. De vraag is dus hoe zich dat ten aanzien van deze materie zal uitkristalliseren. Uit de onzekerheid ten aanzien van de authenticiteit van digitaal bewijs zou mogelijk kunnen volgen dat bewijs steeds minder hard wordt en primair zal dienen als ondersteunend in een bepaald scenario.

Een ander aspect is dat door onszelf gegenereerde digitale informatie straks waarschijnlijk ook makkelijker ter discussie wordt gesteld. In dat opzicht is het Content Authenticity Initiative[541] relevant; hoe neem je reeds bij het maken van het digitale materiaal op transparante wijze echtheidskenmerken op.""

Momenteel is de coördinatie van de benadering van Deepfakes/synthetische media nog primair belegd bij één persoon, maar door de aandacht is er een groeiende community van mensen die ontwikkelingen deelt en binnen het eigen organisatie onderdeel voorlichting geeft. Verder heeft de nieuwe portefeuille Digitale Opsporing het onderwerp geagendeerd voor 2022. Hierbinnen valt ook de ontwikkeling van een toolbox die kan ondersteunen bij het detecteren van (ook niet synthetische) manipulaties. Technische hulpmiddelen kunnen maar ten dele helpen. Uit experimenten/challenges blijkt dat de beste Deepfake-detectietechnieken op dit moment niet meer dan 65% van de Deepfakes kunnen herkennen, dat wil zeggen als dat niet in een gecontroleerde en afgebakende setting gebeurt, maar 'in het wild', zoals op sociale media. "Mijn verwachting is dat de succesratio van deze counter-technologieën nooit significant boven dit percentage uit zal stijgen, onder meer omdat je de detectietools kunt gebruiken om systemen te trainen tot betere deepfakes."

Maar er komen ook steeds ook nieuwe AI technologieën op de markt. Een experiment met een fake-filmpje van Femke Halsema voor de workshops binnen de politie laat zien dat deze, een jaar nadat de betreffende app op de markt kwam, nog steeds niet door de bekende openbare detectoren herkend wordt.[542] Enkele producenten van detectietools gaven ook aan dat dit ook klopt, omdat de gebruikte AI niet onder de definitie van deepfakes valt. Ook daar wreekt de nauwe afbakening van het begrip deepfakes zich dus. Toch is deze reactie verklaarbaar, want als alles synthetisch wordt zal een detector continu aanslaan omdat het (nog?) niet mogelijk is om onderscheid te maken op basis van kwaadwillende intenties.

Intern gaat de politie daarom deels op detectietechnologieën inzetten en vooral ook op het opstellen van goede procedures en normen in afstemming met de keten.

Tot slot zou er ook kunnen worden ingezet op regulering. Dat zou primair op Europees niveau moeten gebeuren, al valt wel op dat Nederland in vergelijking met andere Europese landen een weinig proactieve rol speelt in het aanspreken van Big Tech op bijvoorbeeld privacy schendingen. Nederland zou best iets meer bovenop aanbieders/platformen kunnen zitten qua regulering/aanpak. Qua inhoud lijkt een verbod op Deepfakes/synthetische media niet reëel en bovendien gooi je daarmee het kind met het badwater weg, gezien

de vele positieve toepassingsmogelijkheden. De achterliggende technologie wordt al op zoveel plekken toegepast, zoals in de bedrijfsvoering en dienstverlening van veel organisaties, in de kunst, cultuur, commercie en medische sector. En als je kijkt naar de toepassing door criminelen, dan is wat ze doen al strafbaar...

Het grootste knelpunt zit momenteel bij de aanbieders van malafide toepassingen van synthetische media. Denk daarbij bijvoorbeeld aan de Telegram-bot waarmee je een foto van iemand kan delen, waarna je vervolgens een gefabriceerde naaktfoto van de betreffende persoon terug krijgt. Daar zit nu een veel groter gevaar dan bij de meeste consumentenapps die nu in de appstores voor burgers te downloaden zijn; die apps worden doorgaans voor relatief onschuldige toepassingen ingezet. Een andere mogelijkheid die gesuggereerd wordt is om alles wat geheel of gedeeltelijk synthetisch is gemanipuleerd van een watermerk te voorzien, waardoor transparant wordt in hoeverre het materiaal is gemanipuleerd. Ook daar zie ik geen toekomst in, omdat je in een synthetische wereld dan alles moet waarmerken. Of waar trek je anders de grens, hoeveel manipulatie moet er zijn voordat je het verplicht?

Ook op elke foto die je met je telefoon maakt moet dan zo'n kenmerk komen, helpt dat?

Die 90% is moeilijk te bevatten voor mensen, maar het gaat snel, exponentieel!

10.2.6 Iacobucci

Interview with: **Serena Iacobucci**
Interview by: Bart van der Sloot
Date: 01 07 2021

One of the core problems with deepfake technology is that a clear definition is lacking. When reading the news articles and academic papers on this topic, various non-congruent definitions can be found. In its most basic form, deepfakes refer to putting the face of one person on a picture or video file about another person. Other definitions rely on the technical working behind deepfakes, such as neural networks and GNA networks. Still others refer to manipulations of online content through Artificial Intelligence more in general.

It is also important to distinguish deepfakes from cheapfakes.

Similarly, there is a definition question with respect to one of the use cases of deepfakes, namely when it is used to spread false information. Is it misinformation, disinformation, malinformation, manipulation, etc.?

Positive use cases include bringing deceased persons back to life. For example, you could have artists of ages ago give a tour in a museum, have holocaust victims discuss their experiences or have historical figures teach in school classes.

Dangers include, inter alia, use of deepfakes for non-consensual sexual porn, fraud and misinformation (post-truth era). Although deepfakes obviously help in furthering these malicious goals, many issues have an underlying cause. There is disrespect for women both offline and in particular online; we live in a world were opinions and facts, truths and untruths are increasingly difficult to disentangle; etc. That means that deepfakes are not the problem as such, the problem lies on a deeper, societal level.

"Providing users with a clear definition is one of the first steps to make them willing to engage with a new phenomenon, and this is crucial when we're talking about a phenomenon we want to debunk through education and information. I am afraid that some definitions, together with the fact that they were not clear, were also framed in a negative way - to such an extent that they might have triggered avoidance rather than seeking to familiarize with deepfakes. First, they did not leverage what deepfakes have in common with other manipulated content we are already familiar with and already know how to debunk; Second, they exploited the negative attitudes we might already hold against AI and, also, I believe that the negative attitudes towards an AI is easily spread across all similar AI-generated technologies[543] - so that even the good DF applications have suffered from such negative framing.

I raise the educational/cultural perspectives because, as a matter of fact, concerns regarding deepfakes threats are not balanced. We have a huge (of course, justified and rightful) concern that deepfakes could be a threat for democracies and politics - although, research is showing that political deefpakes are no more effective in deceiving users than other sort of more familiar/ less hyped forms of misinformation[544] and, even when they do impact politicians reputation, the impact is not as worrying as expected[545] - maybe mostly due to the fact that political deepfakes are mostly cheapfakes?.

On the other hand we have the real, ongoing threat, which is already out there as we talk(ed): 96% of deepfakes online represent non-consensual deepfake pornography, most of it portraying women - and as correctly reported by Joseph Cox - "The news acts as a reminder that although in the future political actors may adopt deepfakes for the purposes of disinformation, at the moment their use is squarely in their original, designed purpose: to target and harass." We can blame the technology, for sure - I guess part of this is indeed due to its accessibility and ease-of-use - but still, the applications of the technology are telling us a different story, and I guess it's a cultural one. We can ban, forbit, regulate - but the use of deepfakes for sexual purposes will probably keep existing - even if poorer in terms of quality - if we don't find a way of educating users about what such an appropriation of one's body and image means for the target people. "

Interestingly, there seems to be a normalisation with respect to the receptivity towards deepfakes. Initially, there was a very realistic deepfake of Matteo Renzi, which had a big impact on society and shocked people. Right now, people are not so shocked anymore when they hear something was a deepfake. As a negative, people increasingly start to doubt content. When something offsetting is shown in a video, especially older people tend to doubt its veracity, and belief that it might be a deepfake. As a second negative effect, when people belief a deepfake, this might lead to negative consequences. For example, there was a deepfake about a famous television star, who, judging from a fake video, appeared to have broken a leg after kicking an animal, which led to many death threats of that person.

As a response to deepfakes, regulation can only bring us so far. Essentially, it is not the technique that does harm, but the people that use it for malicious purposes. Hence, the technique should not be curtailed, but the malicious usage of the technology can only be addressed post hoc.

In addition, deepfake detection technology does not look hopeful. Currently, detection programs are only able to pick out part of the deepfakes and most likely, the percentage of deepfakes that can be discovered with the help of technology will be lower in the future.

What is of utmost importance is that experts inform both the general public and policy makes about this new technology and explain how the technology actually works and debunk the myths and fallacies about deepfakes that exist. Currently, much of the debate is uninformed, which leads to thoughtless responses to incidents and hypes. Rather, experts, academics, technology giants and policy makes should join their forces and start an interactive and interdisciplinary debate and cooperate in regulating the technology.

10.2.7 Kirchengast

Interview with: **Tyrone Kirchengast**
Interview by: Bart van der Sloot
Date: 22 06 2021

Deepfake technology can be used to manipulate images, audio and text with the help of Artificial Intelligence. Manipulation of images and other material is obviously not a new phenomenon, but something that has been an problem for many decades if not centuries.

Still, there is a number of significant difference, such as, but not limited to, the ease with which the technology can be deployed, the ease with which material can be distributed and the fact that it can be difficult to impossible to establish whether media is authentic or not. In addition, it may be difficult to establish who the perpetrator behind a deepfake is, particularly where an image is modified by AI and where there is no obvious perpetrator.

Use cases of deepfake technology includes both positive and negative ones. Positive applications include the use of deepfakes for parody/fun, the use of the technology by the movie industry and the use of deepfakes for video-calling purposes. Negative one include porn/sexual deepfakes, deepfakes for political purposes either by citizens or by foreign powers and commercial fraud.

There is also general and specific legislation in place in most jurisdictions to address deepfakes. Non-consensual porn/sexual imagery is regulated, copyright might apply, fraud is obviously criminalised, a regime of civil liability might apply, etc. The problem is that these are discrete statutory instruments. In federal states, the regulatory regime is even more complex. This complexity intensifies the uncertainty with respect to the legitimacy of certain uses of deepfakes and the complexity for victims of deepfakes that want to obtain redress through legal proceedings, because jurisdiction is often shared between states and the federal level. Knowing where to turn for redress can be complicated and confusing.

If governments want to further regulate, the question is where regulatory pressure should be put: on the developers of the software, the app stores and places where people can obtain the technology, on the users of deepfake technology, on the platforms where deepfakes are shown and distributed, etc.? In addition, what makes regulation of deepfakes so complex is the international component. Different jurisdictions apply in different parts of the world, while it is easy to produce and upload deepfakes from

one jurisdiction that are consumed by citizens living in another country. Recourse to private international law and recognition of local court orders across different countries of dissimilar legal tradition may also be required, further impeding access to remedies for individual victims. Such redress is complex, expensive and beyond the reach of most victims.

Education and public awareness campaigns should be part of the solution, but are not a panacea. Technological solutions should also be considered. Watermarking material, introducing authentication systems, counter deepfake technologies and deepfake detection technologies may all be valuable instruments in addressing the problem, but will most likely not be able to provide a full solution.

Social media platforms, in many respects, hold the key to this problem. They should consider their ethical and societal responsibility. They can invest in detection technology and should include clear rules and guidelines in their terms and conditions and sanction users that violate those rules.

Potentially, the government can invest in an ombudsman or e-Commerce or e-Safety Commissioner, that is able to represent citizens, in particular vis-à-vis websites and social media. In general, an individual victim's position is weak when trying to obtain redress and invoke their rights in the online environment and it may be difficult for them to get in contact with moderators of websites, especially where dispersed internationally outside the local system/state under which the complaint arises. Possibly, when contact is laid by a governmental ombudsman, companies will be more inclined to listen. Such an approach should be underpinned by strict community guidelines developed privately by social media companies, which seek to listen to and work with individual victims, to take their complaints seriously, in a timely way. Such guidelines may outline how individual firms may work with government officials or ombudsmen where representations are made. A well-crafted set of community guidelines may obviate any need for further government/ombudsmen intervention, and will help individual victims gain immediate control of manipulated content, potentially limiting further distribution. Establishing workable guidelines should be the first step for any social media firm.

Prohibition of certain types of applications, in particular the sublime deepfakes for non-consensual sexual purposes may be banned. Still, it might be difficult to regulate, as even apps to produce sexual deepfakes may also be used for consensual purposes. In addition, prohibiting apps or service may also backfire; it may make it more interesting or 'cool' to use these technologies.

Finally, judges should be educated with respect to deepfakes. Deepfakes will be used in court rooms, making it more difficult for judges to establish the veracity of material. Judges should be aware of this danger and should be able to send evidence to experts/labs to verify its authenticity. Alternatively, judges may require parties to furnish a certificate that any media tendered in proceedings has not been modified. New jobs are required in this field. An obvious parallel in this respect is the use of DNA material in court rooms, which led to a specific niche for DNA experts and labs that can establish the usability of DNA in court cases.

Potentially, if people are found to have produced fake evidence with the intention of misleading the court, they should be charged and prosecuted under criminal law.

10.2.8 Kwok

Interview with: **Andrei Kwok Onn Jui**
Interview by: Bart van der Sloot
Date: 01 07 2021

Deepfakes are video and/or audio files that are manipulated with the use of Artificial Intelligence (AI). Obviously, the manipulation of media is as old as media themselves and many of the problems we encounter today are similar to those we have already seen with, for example, photoshop. Still, there are certain differences, such as the ease with which manipulated media can be used, the relative low cost for producing deepfakes and the fact that deepfakes are almost impossible to distinguish from the truth.

Right now, there is much attention for the negative use cases of deepfake technology. But the problem is that this focus may have a self-fulling effect. If people only hear about negative use cases, they will either be scared of the technology or will be brought to the attention of the possibility of committing devious acts with the use of this technology. The other way around, if there would be a positive narrative on deepfakes and if there would be many positive best practices for people and businesses to look at, this would influence their behaviour and ensure that more responsible use is made of the technology.

'Deepfake can enhance visitor learning engagement and experience in galleries. Museums worldwide can benefit from this revolutionary convergence between art and artificial intelligence, thus bringing famous artists and their artworks to life. Such experience is possible through the Dalí Lives exhibition in the Dalí museum, St. Petersburg, USA. The museum uses deepfake to recreate an immersive visitor interaction and learning about Salvador Dalí (1904-1989) from the renowned artist himself. Improved deepfake techniques, which overlay high-resolution image blending and image-to-image translation, can transform visitors' experience, thus adding a new dimension of personalized service. With the demand for celebrity images in advertisements and destination marketing, digitally realistic rendition of videos without the celebrity having actually to perform in it has proven possible with significant flexibility. Another example that could be adopted in tourism promotion is Québec's Crown corporation's 50th-anniversary advertisement using deepfake to reminisce viewers of a 1970s version of Bernard Derome, a well-known Canadian news anchor.'[546]

With respect to addressing the potential negative use cases of deepfakes, the following strategies should be considered:

- Reframing the narrative/discourse on deepfakes from negative to positive
- Invest in detection techniques: there will always be a battle between deepfake creation and deepfake detection techniques, but having a deepfake detection filter on your device may in time be just as quintessential as anti-virus software is now
- Social responses: think of the many fact check websites already out there, which could be

used to counter misinformation spread through deepfakes

- Regulation: the state should step in and ensure that beneficial use cases can be pursued, while the malicious ones are curtailed. Regulation should not stifle innovation and creativity, but should ensure that both individual and communal values are safeguarded. In particular, one of the questions is what to do with deepfake technology on the consumer market. For regulators, it is particularly difficult to assess the legitimacy of the millions of deepfakes that can be created by consumers, while the relative value of the use of deepfakes by consumers is low. Most value is to be gained through the use of deepfake technology by commercial parties and industry, and potentially some public sector organisations.
- Social construction of technology perspective: Potential use of a technology will draw in stakeholders who will shape its development. When the stakeholders find beneficial uses of the technology, they will develop it towards a positive direction. Responsibility for the technology development will lead stakeholders towards self-regulation (e.g. establishing code of ethics).

10.2.9 Li

Interview with: **Hao Li**
Interview by: Bart van der Sloot
Date: 30 06 2021

As to the question of definition, AI synthesized Media is a more general term than deepfakes. Deepfakes is for face replacement, whereas AI synthesized media can be for other applications like face reenactment or generating a face.

Deepfake technology is relatively recent and has evolved rapidly since. Most likely, the technology will continue to evolve and there might be radical or qualitative changes, but what is sure is that there will be gradual and quantitative ones: deepfakes will continue to be faster, cheaper, better/higher resolution, less easy to detect because the technique leaves behind less artefacts and ever more accessible for laypersons and consumers, e.g. via apps.

The growth in terms of the use of this technology is to be expected in the consumer market: people finding it cool to experiment with video and face-swap, to virtually restyle their own faces and looks or to engage inf forms of parody. The entertainment industry also use the technology for de-aging people or to have a virtual double to the stunts for the actor. The core technology can be deployed for creating avatars and realistic digital humans, which can be used, inter alia, in video games. In addition, it can be deployed for, for example, having all non-English movies translated and having the actors speak the lines in English, so as to make those movies accessible to the whole world.

Obviously, there will also be a rise in negative use cases, such as, but not limited to, the creation of non-consensual sexual images, the use of deepfake technology for misinformation, influencing democratic elections and for committing fraud and deceit in cyberspace.

The problem in addressing the negative aspects of this technology is threefold:

Deepfake detection technologies will always run behind in the arms race between deepfake creation and deepfake detection. As deepfake

creation technology will continue to be more and more accurate and leave less and less traces behind, there will come a point were deepfake detection techniques may lose the battle. In addition, it is important to stress that deepfake technology can also be deployed in order to make real images/videos look fake, for example by leaving traces of manipulation behind, without the content in fact being manipulated. From a practical point of view, the best way to detect deepfakes will be contextual information. Is this something this person would have said? Would he/she speak this way or walk that way? Does this piece of information fit together with other pieces of information – does it make sense in light of other sources that are known to be authentic?

Education and awareness might have some value, as a number of people are unaware of the dangers of deepfakes. This might help prevent them from believing false information that is inauthentic (though presumably only marginally so). However, with respect to the production of deepfakes, those that procedure deepfakes for malicious purposes – either non-consensual sexual imagery, misinformation or fraud – very well know that they are engaging in unethical behaviour and find this new technology a useful instrument in their intentions. It is unsure to what extent public awareness campaigns might be effective.

Finally, regulation should not go so far that it outright prohibits deepfake technology. But regulation and action are absolutely essential. Malicious deepfakes are in that sense similar to spam – if you would not curb spam, e-mail as such would become practically useless. Similarly, the internet and social networks need to be relatively free of malicious deepfakes in order to continue to be a valuable and free space and to retain the value they have in our current day society.

10.2.10 Maddocks

Interview with: **Sophie Maddocks**
Interview by: Bart van der Sloot
Interview date: 17 06 2021

Deepfakes are primarily used for pornographic or sexualised content. Estimates go as high as 93% of the deepfakes being pornographic in nature. This most likely will remain the most important use case of this technology. Thus, if one wants to give a definition of Deepfake, the pornographic use case should be included. Something like: 'Deepfake: content created in whole or in part by Artificial Intelligence....., in particular used for the creation of pornographic material'.

As such, deepfakes fit in a broader trend on the internet to sexualise women. They are a networked object and are used within a certain online environment. By far most sexual deepfakes are non-consensual, either as revenge porn or directed at celebrities. In addition, use is often made of material about sex workers. 'The creation and distribution of private sexual content without consent is also a threat to the autonomy of porn performers and sex workers. Deep fake porn is the newest incarnation of NCP [Non Consensual Porn] and it disproportionately targets porn performers. Using face-swapping technology, deep fakes create realistic fake porn by editing victims' faces into existing pornographic content. Porn performers have reported feeling "eviscerated" when their content is erased, edited, and recirculated as deep fake porn.'[547]

There is too much focus in the debate on political use cases, which represent a small minority of deepfake only. Obviously, they can do grave harm, if they are used to undermine democratic processes or the rule of law. But currently and most likely in the future, the harm caused by pornographic deepfakes by far outweighs the harm caused by political deepfakes. Sexual harms may include humiliation and may in extremo lead to honour killings or suicide.[548] In addition, they might normalise certain behaviour that is currently deemed in appropriate.

For sexual fakes, often, use is made of 'cheapfakes', that is, technique that is not as advanced as the technology deployed in the film industry or by states to disrupt democratic processes. However, users of sexual fakes are usually not bothered by the low fidelity quality of the output; the mind, in a certain sense, does the rest.

Obviously, both deep and cheapfakes raise evidentiary questions. These questions as such are not new, they resemble many of the difficulties encountered both in the offline and the online world. However, as *Britt Paris and Joan Donovan stress:* '"New media technologies do not inherently change how evidence works in society. What they do is provide new opportunities for the negotiation of expertise, and therefore power."[549] Obviously, deep and cheapfakes may make arriving at a conviction more difficult, but in the online context, the chances of being charged and convicted are already quite low.

As a potentially dangerous new use case of deepfakes one could point to the 'live deepfakes', which seems to have an audience. As a morally more complicated use case, one can refer to the deepfakes being deployed for the treatment or help of paedophiles, with the hope that the imagery will deter them from effectuating their desires on real children.

Regulation is difficult in this context. An outright prohibition seems difficult to enforce and moreover, it might disempower women. There are also women who want to use deepfake technology, either for professional or for personal usage. In addition, non-consensual sexual deep and cheapfakes are currently already prohibited in most jurisdictions, as consent is required. That is why more time and energy needs to be invested in education and awareness of the (online) public. The victim's voice should be heard more often.[550]

Finally, platforms should take their responsibility in assessing to what extent content shown on their network adheres to all legal, societal and ethical standards and requirements. Obligations and regulatory pressure should be imposed throughout the chain.

10.2.11 Perot

Interview with: **Emma Perot**
Interview by: Bart van der Sloot
& Yvette Wagensveld
Date: 14 06 2021

Deepfake is a form of digital manipulation. Deepfake technology can be used to create a video or audio file that is seemingly authentic. This is what makes it potentially dangerous.

The technology is and most likely will continue to be deployed primarily for pornographic usage. Obviously, there are also positive use cases,

such as in the film industry or just silly deepfakes produced by citizens for parody. A particular development will be the production of deepfakes of people who are deceased. These can be used for educational purposes, such as having a historical figure teach a lesson to school children or bringing family members that have passed away back to life.

There will most likely be an increase in political deepfakes and deepfakes for intentional misinformation as well, as countries are investing heavily in techniques to disrupt each other's democratic processes. But disruptions could be conducted by a group that is not state sanctioned, or even groups within the country rather than third party states as well. Most likely, it will be primarily the negative use cases that will grow in the future, while positive use cases will be restricted to certain domains.

Deepfake technology fits into a wider trend of manipulation of content. Misinformation in the online environment is endemic. People often confuse opinion with fact. This is what you see during the pandemic and the misinformation that is distributed about both the virus and vaccines. Deepfakes can incite confusion and misconception, and can potentially cause grave harm. Small manipulations are becoming the standard, think of filters to equalise skin or adjust skin tone. This may create an environment in which it is increasingly difficult to establish what is true and what is not.

This also may make it more difficult to enforce your rights, to enforce the current legal regime, as it might be difficult to establish what is true and what is not, who manipulated content, how and why, and who could be held accountable for it. Legal processes, even if you have a valid claim, can be long and costly.

In an attempt to curtail deepfake technology, counter technologies, such as deepfake detection technologies, can only bring us so far. They can only filter out part of the deepfakes and can often not establish with certainty which part of the video or audio file has been manipulated.

There is currently regulation in place, such as human rights legislation, privacy principles, intellectual property rights and persona rights. With respect to the latter: 'In both the UK and USA, deepfakes are likely to be addressed by current law. In the UK, passing off, while more complex than the right of publicity, could be helpful to address deepfakes where they are used in a way that conveys endorsement. For non-endorsement uses, defamation might be grounds for legal action. In the USA, the right of publicity can be deployed. However, these causes of action inadequately address the potential of deepfakes to sway public opinion, damage the reputation of the individual depicted, and impinge on privacy and dignity of the individual if used in pornography, for example. It may be challenging to track down the uploader, and the delay between commencing legal action and receiving an injunction can be too late if the deepfake goes viral. The potential velocity and amplification of deepfakes on the web allows for nefarious uses of this relatively new technology.'[551]

Currently regulation is clear on certain use cases. Child pornography; clear case. Revenge porn and other forms of non-consensual pornography is often easy to categorise. Intellectual property law can be applied to commercial use cases, but should be mainly limited to that. There is a

risk in overstretching the current legal regime to apply to use cases the rules were not designed for. Rather, new regulation should be considered.

What makes further regulation complicated is the fact that people also have the freedom of speech. In addition, in most jurisdictions, privacy rights of celebrities and public figures are limited. But especially when deepfakes are used in the political process, to disrupt democratic processes, strict regulation and a proactive attitude of governmental authorities is required. In addition, manipulation of evidence in court rooms could be explicitly criminalised.

Awareness campaigns could be helpful, as people should be more aware about the dangers of misinformation in general. Social media platforms should cooperate as well, as they hold a big part of the key in successfully addressing the issue of misinformation and deepfakes.

10.2.12 Stevens

Interview met: **Lonneke Stevens**
Interview door: Bart van der Sloot
& Yvette Wagensveld
Datum: 23 06 2021

De veronderstelling van de rechter bij aangeleverd bewijs is dat het materiaal betrouwbaar is, tenzij de verdachte/de verdediging het tegendeel beweert. Daarbij moet er wel een goed verhaal worden gegeven. Je kunt niet zomaar roepen: dit klopt niet/het is nep. Dan zal je moeten uitleggen waarom het aannemelijk is dat bewijs niet deugdelijk is. Bijvoorbeeld, op deze video ben ik te zien op moment x, waarbij ik gedraging y vertoon, maar op moment x was ik helemaal niet daar waar ik lijk te zijn in de video.

Het systeem voor de beoordeling van authenticiteit van bewijs is momenteel vrij los. Het is in ultimo aan de recht om hier een oordeel over te vellen.

Vermoedelijk zullen er op termijn deskundigen in het register worden bijgeschreven die meer expertise hebben op het punt van deepfake technologie en de mogelijkheid om sporen van het gebruik van deze technologie te achterhalen in aangeleverd bewijs.

Videomateriaal is momenteel nog geen wettig bewijsmiddel, maar dat verandert door de modernisering van het wetboek van strafvordering. Het wordt nu gebruikt via het bewijsmiddel van de eigen waarneming van de rechter, en in de vorm van een proces-verbaal van politie waarin wordt beschreven wat op het beeld te zien is. De erkenning van wettig bewijsmiddel is ook een belangrijke stap naar het kunnen stellen van betrouwbaarheidseisen aan videomateriaal.

Er zijn wel bewijsmiddelen die zo ondeugdelijk zijn dat die simpelweg niet worden toegelaten. Een voorbeeld daarvan is het gebruik van een leugendetector. Maar als deepfake technologie kan worden ingezet om zowel beeld, audio als tekst te manipuleren, dan is het onwaarschijnlijk dat als stelregel wordt aangenomen dat al dit type materiaal niet in de rechtszaal mag worden aangevoerd omdat het gevaar is dat ze zijn gemanipuleerd; dan blijft er immers weinig over.

Wellicht kan de discussie over deepfakes in die zin een professionaliseringslag betekenen. Het vermoeden is momenteel dat er weinig onbetrouwbaar/vervalst bewijsmateriaal bij de rechter wordt aangeleverd, maar zeker is dat

uiteraard niet, omdat er niet een standaardcheck wordt gedaan op dit punt. Door meer in te zetten op controlemechanismen kan op het punt van deugdelijkheid en betrouwbarheid van bewijs meer in algemene zin winst worden behaald.

Voor het werken met waarschijnlijkheidspercentages – de kans dat deze video authentiek is, is bv 76% - zal veel worden gekeken naar steunbewijs. Het gaat om het geheel aan feiten en omstandigheden waaruit een uiteindelijke overtuiging van de rechter volgt.

Ook kan worden gekeken naar hoe rechters omgaan met DNA materiaal. Daarbij heb je gecertificeerde bureaus die kunnen vaststellen hoe betrouwbaar materiaal is en wat de kans is dat het DNA van de verdachte is. Rechters blijken daar in de praktijk moeite mee te hebben. Daarom wordt wel langer gepleit voor meer en betere opleiding van rechters, zodat zij meer achtergrond kennis krijgen van statistiek, methodiek en bepaalde technische kennis die vereist is om de waarachtigheid en relevantie van bewijsmateriaal goed op waarde te kunnen schatten.

10.2.13 Yaldin

Interview with: **Aya Yadlin**
Interview by: Bart van der Sloot
Date: 15 06 2021

The debate about deepfakes resembles the broader debate about technology, which tends to take place in extremes. Tech-utopians will claim that technology is going to save us, make the world a better, safer and more efficient place, while tech-dystopian will claim that technology is not working and that 'AI/the robot/etc. is going to kill us'. Journalist often use this frame, both because these stances do not require detailed knowledge of technology and because alarmist appeals attract a big audience more easily than nuanced output. Journalists themselves generally do not have a technological background and thus use a cultural lens and cultural concepts to explain and understand technology.

This same trend is visible within the discussion on deepfakes, in particular the dystopian side. Alarm is raised over all kinds of real and hypothetical use cases, while little information is provided on how the technology works, why it might also benefit society, etc. This is problematic because it leads to an uninformed debate, both in the media and as a consequence, with the population at large and regulators. There is a double lost in translation. Journalists trying to grasp new technologies understand it only partially and the readers only understands parts of the news item, or make up their own story about the technology.

Deepfake was a term that was originally used to describe a small use case, primarily (but not only) referring to porn based fake images. Gradually, it came to stand for a bigger phenomenon, namely that of synthetic media, created with the help of AI.

Deepfakes are not a standalone issue, they fit in a general environment. Misinformation is a general problem; deepfakes can obviously help in spreading misinformation, but the underlying trend is already ongoing. The same applies to sexual/pornographic deepfakes. There is a general problem with respect to the treatment of and respect for women, both offline and in particular online. Deepfakes can and are also

used for political purposes, but again, they can be particularly powerful when they fit into a general trend of societal reality. For example, in Israel, they might be used to further the tension between the Arab and Jewish population.

The extent to which media represent reality has always been a discussion. From Plato onward, people have wondered to what extent our ideas about reality are a correct representation. Every new medium has spiralled this same debate, e.g. with the rise of the internet, the same discussion existed. Is this a new reality, will it be a parallel reality, will it replace our reality, etc.? With the introduction of the printing press, there were concerns over the privatisation of media and the truth. That is not to say that nothing new is happening with deepfakes. There is a cultural shift and visual turn ongoing, people rely more and more on visual images for their news consumption and understanding of reality. In addition, deepfakes can be so real that it is difficult to asses their authenticity. While texts describe reality, deepfakes mimic reality.

Positive use cases are also evident. An example might be David Beckham who appears to be speaking all languages in the world when addressing the public concerning a public/good cause. Other examples are uses of deepfake technology in the film industry and for business applications and video calling.

Addressing the negative consequences of deepfakes may be difficult, but one more general problem is the diversification of the information landscape and the lack of editorial control by various platforms and gatekeepers. In addition, if these gatekeepers do exert control, they are biased and have their own (commercial) interests. That is why a bigger role for Public Service Media may be advocated. It is generally thought that PSM have lost their significance in the online environment, but the contrary is true. During the pandemic, citizens have turned to PSM again for reliable, objective and neutral news and PSM are leading in many countries in new media outputs, such as podcasts. In Israel, there is a large public podcast platform. An option might be to invest more in PSM and further international collaborations, so that PSM can function as gatekeepers and assess the authenticity of news and information.

A second option is to invest more in audience literacy. People know very little about the possibility to manipulate media, video, audio and text. In particular, when they see a video, they tend to believe its authenticity and when communicated to them by a friend or shared with them in an online community they are part of, they tend to trust the source. This is where awareness and education might help. People should be more aware of the potential of misinformation and should learn to verify content they consume.

Regulation is already in place and many deepfakes are already prohibited, such as those concerning non-consensual pornographic deepfakes. The problem concerns primarily the enforcement of the regulation in the cross-border online context. A prohibition of or licencing deepfake technology for the consumer market might be an option, though the question is on which entity/where in the chain this obligation will be put: tech developers, platforms, app-stores, consumers, etc.

10.2.14 Westerlund

Interview with: **Mika Westerlund**
Interview by: Bart van der Sloot
Date: 18 06 2021

The definition of deepfakes has changed over time. In the beginning, the concept was used to refer to a small number of use cases. Gradually, it came to stand for any content that is manipulated with the use of Artificial Intelligence/Machine Learning. Deepfake technology can be used to manipulate images, sounds, text and video.

An important new addition is the manipulation of satellite images and signals. Satellite images are in the hands of a small number of parties. When one party manipulates the images, another party may produce the real/authentic images, but for the larger audience, it will still be difficult to establish which version is correct.

The most important use cases, both negative and positive, are set out in the article from 2019.[552] Political use cases have been limited so far, but most likely, the use of deepfakes for political purposes will grow in the coming years. Porn has been the biggest use case and most likely will continue to be. The next big revolution in deepfake technology will presumably concern the business market, the use of deepfakes by professional organisations and for commercial usage. The second growth in the use of deepfake technology will most likely take place in the consumer market, in particular the use of the technology for parody and the creation of funny output.

Deepfake technology fits into a general trend of manipulating media, of the 'post-truth' era, in which opinion and fact are increasingly difficult to distinguish, and in the creation of misinformation. In that sense, nothing is radically new, although of course the ease with which material can be manipulated and distributed and the difficulty of establishing what is authentic and what is not is essentially different for deepfakes than for previous technologies.

It is difficult to ban deepfake technology, even if a government would attempt to do so, both because the international/cross-border environment and the fact that many deepfake software is available in open source. Thus, it seems hardly possible to contain the use of this technology.

Governmental organisations and in particular intelligence agencies are investing heavily in deepfake detection technology, but it is questionable how effective this technology will be. Mostly likely, deepfake technology will be so advanced and deepfakes will be so difficult to detect in a few years' time, that most deepfakes will not be spotted by deepfake detection technology.

Already now, experts have difficulty in establishing whether content is deepfake or not. An example is the famous case of a mother who was believed to have created deepfakes about the team mates of her daughter on a cheerleading team. She was believed to have produced negative content about them and was brought to court. But right now, experts in the legal proceedings have difficulty in establishing whether the content is fake and if so, which part of the content is fake and which part has been fabricated by the mother.

This example also raises another important concern deepfakes spiral, namely that there

will be many problems with evidence in court rooms. It is so easy for citizens to use deepfake technology and to produce material that supports their case/point of view that this will most likely be an important use case in the near future.

The most realistic scenario for containing deepfakes seems not so much the production of deepfakes, but their distribution. That is why platforms and social media should play an important role. They already have duties of care to prevent and counter illegal content and prevent harm, but these may be extended where necessary.

Finally, it would be wise to invest in education/ public awareness. People generally know little about deepfake technology and think they are good in detecting whether a piece of information/ video/audio presented to them is authentic or not, while the contrary is true. What makes public awareness difficult is that people not always care whether something is true or not.

10.2.15 Whyte

Interview with: **Christopher Whyte**
Interview by: Bart van der Sloot
& Yvette Wagensveld
Date: 23 06 2021

Deepfake refers to any material that is manipulated or generated with the use of AI. This can be audio, video, text and signals. The entertainment industry had some technology that resembles what deepfake technology can do now, but in essence, the technology is relatively new. It is only about four years old. The technology has already developed significantly in those years and most likely, adversarial models and machine learning applications will continue to develop rapidly the coming years.

The estimate by some that in a small number of years, more than 90% of the online material will be synthetic seems realistic. There are so many filters and small manipulations built into both software and hardware, that the content (pictures, video, audio) produced by consumers often is partially manipulated. Deepfake technology fits into a general environment in which misinformation and distrust in the media and the government is already high. That is why people may be receptive to fake content. Deepfakes may lead to an astroturfing of society.

Deepfakes are part of the problem only, shallow fakes, the low fidelity technology, may have a bigger impact, as this technology will be in the hands of millions or billions of people, which will make it hard to retain a sense of shared reality. Deepfake technology will be in the hands of a small number of players, in particular the entertainment industry and governmental organisations. Governments such as China, Iran and Russia deploy the technology to destabilise Western democracies. Importantly, the 'battlefield' of deepfake technology will not only be on western soil, but also in second and third world countries, that will try to influence the political establishment and the popular opinion there.

The real problem to come will not be individual deep- or cheapfakes, but the ease with which a whole ecosystem of false information can be created. A fake video, but also fake twitter accounts that link to the video, fake accounts on discussion forums that discuss the content of the video, fake websites that host the video

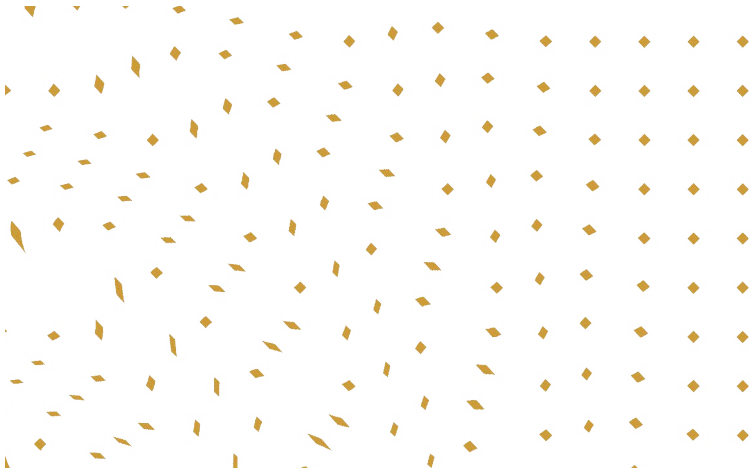
and produce fake news reports on that what is shown in the video, fake Instagram accounts that produce memes of the fake video, etc. It will be very hard to pierce through a multi-layered environment of deception.

Technology may be part of the solution. Digital signatures, public-private key encryption, authentication systems, etc. In addition, there is deepfake detection technology. However, this is most likely only part of the solution. Detection programs are unable to detect all deepfakes and it will be difficult to impossible to ensure that all online content is authenticated via a secure and trustworthy system.

Regulation of deep- and cheapfakes is difficult, because it is difficult to say what is a deep and what is a cheap fake, what are legitimate and what are illegitimate uses, which jurisdiction prevails, that of the producer of the deepfake or that of the person being depicted. In addition, there is a black market for this type of technology. Prohibiting the technology might backfire, as it might have the effect that technology becomes more cool. Finally, at least in the U.S., but in many other jurisdictions as well, the freedom of speech protects the use of this technology by citizens to a large extent.

The most important question will concern the use of deepfake technology by foreign governments. Plain old diplomacy might work. But it should also be assessed how it can be ensured that governments that deploy these technologies on foreign soil can be sanctioned. Finally, it should be assessed how technology companies that develop deepfake technology and sell it, might be sanctioned if they sell it to malicious parties.

The next big moment from an US perspective will be the elections in 2024 and/or 2026. What will the influence of deepfakes be and how will the US respond?



Voetnoten hoofdstuk 10

- 414 Senior research fellow, Tilburg Institute for Law, Technology, and Society, Tilburg University Law School, the Netherlands. s.b.zhao@tilburguniversity.edu.
- 415 Zeyi Yang, *China's viral DeepFakes spook Beijing,* Protocol — The people, power and politics of tech (2021), https://www.protocol.com/china/chinese-deepfakes-regulators-alibaba-tencent (last visited May 9, 2021).
- 416 Following EU's new research report on Deepfakes, Deepfakes are defined as " deepfakes are defined as manipulated or synthetic audio or visual media that seem authentic, which feature people that appear to say or do something they never said or did, produced using artificial intelligence techniques, including machine learning and deep learning." See: XX, p. 1.
- 417 Global times, *Chinese regulator summons major internet platforms for talks on content authenticity - Global Times, Global Times* (Mar. 18, 20211), https://www.globaltimes.cn/page/202103/1218762.shtml (last visited May 9, 2021).
- 418 Emma Woollacott, *China Bans Deepfakes In New Content Crackdown,* Forbes , https://www.forbes.com/sites/emmawoollacott/2019/11/30/china-bans-deepfakes-in-new-content-crackdown/ (last visited May 9, 2021).
- 419 Global times, supra note 417.

- 420 Yu Yuan Qian, Tianjin's legislation forbidding misuse of facial recognition technology in social credit system: not a clear cut rule, Jingji Guancha Newspaper (in Chinese), 3 Dec. 2020, at: http://www.eeo.com.cn/2020/1203/441394.shtml.
- 421 Mainly: photo and video graphic deepfake technology, specific graphical deepfake techniques, voice cloning technology, text synthesis technology, and detection software. See the most recent EU draft report, "Tackling Deepfakes in the new AI legislative Framework and Beyond", Chapter 3.
- 422 See: Tablet B.1 Examples of personal sensitive data, English version at: https://www.dataguidance.com/legal-research/standard-gbt-35273-2020-information-security
- 423 Italicized by the author. See: China's Draft Personal Information Protection Law (English version), at: https://www.newamerica.org/cybersecurity-initiative/digichina/blog/chinas-draft-personal-information-protection-law-full-translation/?utm_campaign=Marketing_Cloud&utm_medium=email&utm_source=Standalone+Call+for+Comments+Personal+Information+-+10.27.20&%20utm_content=https%3a%2f%2fwww.newamerica.org%2fcybersecurity-initiative%2fdigichina%2fblog%2fchinas-draft-personal-information-protection-law-full-translation%2f
- 424 Art. 76 (1), English translation at: Translation: Cybersecurity Law of the People's Republic of China (Effective June 1, 2017), , New America , http://newamerica.org/cybersecurity-initiative/digichina/blog/translation-cybersecurity-law-peoples-republic-china/ (last visited Jun 10, 2021).
- 425 Xu Guodong, Liu Qiang & Liu Haibin, *China's First Facial Recognition Case,* - China Justice Observer (2021), https://www.chinajusticeobserver.com/a/china-s-first-facial-recognition-case (last visited Jun 28, 2021).
- 426 See a detailed discussion of this Regulation at Section 4.2.
- 427 Woollacott, *supra* note 418.
- 428 Huixiang Xing, Regulating Facial Recognition Technology, Comparative law Study (in Chinese), May, 2020, at: https://www.163.com/dy/article/FP3O1NJU0530W1MT.html
- 429 Global Legal Monitor, *China: National People's Congress Adopts New Civil Code* (2020), //www.loc.gov/law/foreign-news/article/china-national-peoples-congress-adopts-new-civil-code/ (last visited May 10, 2021).
- 430 English version, at: https://www.cecc.gov/resources/legal-provisions/criminal-law-of-the-peoples-republic-of-china#2%20Chapter%20IV
- 431 Draft in Chinese, at: https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwiezbuVpbDxAhWIERQKHXywAaEQFjAAegQIBBAD&url=https%3A%2F%2Fwww.dataguidance.com%2Fsites%2Fdefault%2Ffiles%2Fchina_draft_personal_data_law.pdf&usg=AOvVaw3kGuQe2aVVIutfbdHb_TYo
- 432 Personal Information Protection Act (draft) (in Chinese), at: https://www.secrss.com/articles/26427
- 433 English version, at: https://www.newamerica.org/cybersecurity-initiative/digichina/blog/translation-cybersecurity-law-peoples-republic-china/
- 434 Chinese version, at: https://www.secrss.com/articles/17713
- 435 Xu Jiangzeng, Facial Recognition Technology: rising risks and regulation (in Chinese), at: https://new.qq.com/omn/20210406/20210406A048XB00.html
- 436 Copyright Law of People's Republic of China (Chinese, 2020), at: https://mp.weixin.qq.com/s?__biz=MzI4NzcyNTMyNg==&mid=2247485171&idx=2&sn=8c1941b95070e7c572cdd18a31998b17&chksm=ebc80492dcbf8d84b15f28e28fe4325d483780d57f8a662b512035d997cdfb13562bb38892e8&scene=21#wechat_redirect
- 437 This law is still valid, but obsolete due to the rise of the ICt technology. For a Chinese version, available at: http://www.gov.cn/gongbao/content/2016/content_5139387.htm
- 438 Rule of Law in China: strategic plan 2020-2025

(in Chinese), available at: http://www.npc.gov.cn/npc/c30834/202101/571add54b21b40b8a5ab710acb194a8c.shtml

- 439 Full Translation: China's "New Generation Artificial Intelligence Development Plan" (2017), New America , at: http://newamerica.org/cybersecurity-initiative/digichina/blog/full-translation-chinas-new-generation-artificial-intelligence-development-plan-2017/ (last visited Jun 10, 2021).
- 440 Ibid., p. 25.
- 441 Ibid, p. 26.
- 442 Ibid, p. 26.
- 443 Danit Gal, *Mapping official AI ethics discussions in China*, nesta (2020), https://www.nesta.org.uk/report/chinas-approach-to-ai-ethics/mapping-official-ai-ethics-discussions-china/ (last visited Jun 10, 2021).
- 444 For an English version, see: Chinese Expert Group Offers "Governance Principles" for "Responsible AI," , New America , http://newamerica.org/cybersecurity-initiative/digichina/blog/translation-chinese-expert-group-offers-governance-principles-responsible-ai/ (last visited Jun 10, 2021).
- 445 Gal, *supra* note 443.
- 446 Chinese text, at: https://www.secrss.com/articles/11444
- 447 Chinese version, at: http://www.gov.cn/gongbao/content/2000/content_60531.htm
- 448 Chinese version, at: http://www.cac.gov.cn/2017-05/02/c_1120902760.htm
- 449 Chinese version, at: http://www.gov.cn/flfg/2011-03/21/content_1828568.htm
- 450 Chinese version, at: http://www.lawyers.org.cn/info/a39e762b7c8d495bad8499efb1fa4376
- 451 English version, at: http://www.lawinfochina.com/display.aspx?lib=law&id=6582
- 452 Chinese version, at: https://www.waizi.org.cn/doc/66064.html
- 453 Chinese version, at: https://baike.baidu.com/item/%E7%BD%91%E7%BB%9C%E8%A7%86%E5%90%AC%E8%8A%82%E7%9B%AE%E5%86%85%E5%AE%B9%E5%AE%A1%E6%A0%B8%E9%80%9A%E5%88%99/21508108?noadapt=1
- 454 Chinese version, at: http://gbdsj.gd.gov.cn/zwgk/zwwgk/jcgk/content/post_2274724.html
- 455 Chinese version, at: http://www.elawcn.com/copyright/2020/1219/727.html
- 456 Chinese version, at: http://politics.people.com.cn/n1/2019/0110/c1001-30513562.html
- 457 Chinese version, at: http://politics.people.com.cn/n1/2019/0110/c1001-30513562.html
- 458 Chinese version, at; http://wwhttp://www.lawinfochina.com/display.aspx?lib=law&id=6582 w.nrta.gov.cn/art/2019/11/29/art_113_48908.html
- 459 Other industrial standard policies in a similar vein cover security standards of FRT data (2020-5157 - T-321), FRT in security surveillance systems (GB/T 31488-2015), and FRT in gate or entrance control (GA/T 1093-2013). Chinese policies available at: https://www.bi6.org/2020/05/zhongbang____zhongguoshougeyuanchengrenlianshibiejianbieshenfenxitongguojiabiaozhunfabu_19488.html
- 460 Available at: https://www.chinesestandard.net/PDF.aspx/GBT38671-2020 (English and Chinese) ; an official Chinese version available at: http://www.gb688.cn/bzgk/gb/newGbInfo?hcno=C84D5EA6AC99608C8B9EE8522050B094
- 461 Information Technology-Security Techniques -Biometric information protection. (Chinese version), at: http://std.samr.gov.cn/gb/search/gbDetailed?id=7234D8342BD46252E05397BE0A0A000B
- 462 Interpretation of the new law on personal financial information protection: "The Technical Rules on Personal Financial Information" (in Chinese), available at: https://www.secrss.com/articles/17921
- 463 Ibid.
- 464 https://law.unimelb.edu.au/about/staff/andrew-roberts
- 465 'A viral video that appeared to show Obama calling Trump a 'dips---' shows a disturbing new trend called 'Deepfake'', Business Insider Australia, April 18 2018,

https://www.businessinsider.com.au/obama-deepfake-video-insulting-trump-2018-4?r; 'Weaponised Deep Fakes', Australia Strategic Policy Institute, 29 April 2020, https://www.aspi.org.au/report/weaponised-deep-fakes

- 466 Belgian socialist party circulates 'deep fake' Donald Trump video, Politico, 21 May 2018; *https://www.politico.eu/article/spa-donald-trump-belgium-paris-climate-agreement-belgian-socialist-party-circulates-deep-fake-trump-video/*
- 467 'This Deepfake of Mark Zuckerberg Tests Facebook's Fake Video Policies', Vice, 12th June 2019, *https://www.vice.com/en/article/ywyxex/deep-fake-of-mark-zuckerberg-facebook-fake-video-policy*
- 468 See, for example, B. Armesto-Larson, 'Non-Consensual Pornography: Criminal Law Solutions to a Worldwide Problem,', (2020) 21 Oregon Review of International Law 177. A. Schein, 'When Sharing is Not Caring: Creating an Effective Criminal Framework Free from Specific Intent Provisionns to Better Achieve Justice for Victims of Revenge Pornography', (2019) 40 Cardozo Law Review 1953; H. Hayden, 'Rewriting Arizona's Revenge Porn Statute to Fill the Gap in Sex Crime Punishment', (2019) 51 Arizona State Law Journal 397; R. Rosenberg and H. Dancig-Rosenberg, 'Reconcpetualzing Revenge Porn', (2021) 63 Arizona Law Review 199, who suggest that non-consensual pornography has become an epidemic in the United States. On the use of technology to create 'morphed' child pornography, see C. Beacham, 'Metamorphosis: Changing Oklahoma Law to Protect Children from Morphed Child Pornography', (2020) 55 Tulsa Law Review 311.
- 469 'Most Deepfake Are Porn, and They're Multiplying Fast: Researchers worry that doctored videos may disrupt the 2020 election, but a new report finds that 96 percent of Deepfake are pornographic', Wired, 10 July 2019, *https://www.wired.com/story/most-Deepfake-porn-multiplying-fast/*
- 470 C. Gosse and J. Burkell, 'Politics and Porn: How News Media Characterizes Problems Presented by Deepfake', 2020) 37 Critical Studies in Media Communications 497.
- 471 See Report of the Select Committee on Intelligence on Russian Active Measures Campaigns and Interference in the 2016 U.S. Election, Senate Report 116-290,, Volume 2: Russia's Use of Social Media with Additional Views. Available at *https://www.intelligence.senate.gov/sites/default/files/documents/Report_Volume2.pdf [Accessed 3 June 2021].*
- 472 Yvette Clarke, *Deepfake Will Influence the 2020 Election – and Our Economy, and Our Prison System, Quartz: Ideas, July 20, 2019. Available at https://qz.com/1660737/Deepfake-will-influence-the-2020-election/ [Accessed 7 Jun 2021].*
- 473 B. Chesney and D. Citron, 'Deepfake: A Looming Challenge for Privacy, Democracy and National Security', 2019) 107 California Law Review 1753.
- 474 Ibid, 1994
- 475 J. Ice, 'Defamatory Political Deepfake and the First Amendment', (2019) 70 Case Western Law Review 417, 431.
- 476 Chesney and Citron, above n.?, 1801-1802.
- 477 A. Gieseke, '"The New Weapon of Choice": Law's Current Inability to Properly Address Deepfake Pornography', (2020) 73 Vanderbilt Law Review 1479.
- 478 Ibid, , 1501-1502.
- 479 H.R. 3230 – 116th Congress (2019-2020). Text and legislative history available at *https://www.congress.gov/bill/116th-congress/house-bill/3230.* [Accessed 11 June 2021]
- 480 M. J. Blitz, 'Lies, Line Drawing, and (Deep) Fake News, (2018) 72 Oklahoma Law Review 59, 62.
- 481 567 U.S. 709 (2012).
- 482 Alvarez, at 722.
- 483 See Chesney and Citron, 1789, citing *Thomas v Collins,* 323 U.S. 516 (1945) and *Abrams v United States,* 250 U.S. 616 ((1919).
- 484 Chesney and Citron, at 1779.
- 485 Ibid, 103.
- 486 Ibid, 111. See also Chesney and Citron, above n.???, at 1778.
- 487 Ref to Citron on the possible consequences of deepfake pornography
- 488 Texas Election Code, Title 15, Regulating Political Funds and Campaigns, Chapter 255, Regulating Political Advertising and Campaign Communications, Section

255.004(d) – True Source of Communication (Texas Senate Bill 751).

- 489 Texas Senate Bill 751 (Section 1, 255.004(e) Election Code).
- 490 Assembly Bill No.602 - Adds new section - Section 1708.86 - to the Californian Civil Code. In force October 3, 2019.
- 491 Code of Virginia, Title 18, § 18.2-386.2 - Unlawful dissemination or sale of images of another
- 492 Senate Bill No.5959-D, which inserted new sections 50-f and 52-c into the NY Civ Rights L (2014)
- 493 Senate Bill No.5959-D, which inserted new sections 50-f and 52-c into the NY Civ Rights L (2014)
- 494 H.R. 3230 – 116th Congress (2019-2020). Text and legislative history available at *https://www.congress.gov/bill/116th-congress/house-bill/3230. [Accessed 11 June 2021]*
- 495 SB 751, Bill Analysis, Senate Research Center, 29 March 2019. Available at *https://lrl.texas.gov/scanned/srcBillAnalyses/86-0/SB751INT.PDF* [Accessed 12 Jun 2021.
- 496 California Elections Code – Division 20 Election Campaigns – Chapter 1 Endorsement of Candidates - § 20010.
- 497 HR 1 – 117th Congress (2020-2021). Introduced 4 January 2021. Text and legislative history available at https://www.congress.gov/bill/117th-congress/house-bill/1/actions. [Accessed 12 June 2021]
- 498 The offence also covers voyeuristic recording using covert recording devices and 'up-skirting'.
- 499 Virginia Code § 18.2-386.1
- 500 Virginia Code § 18.2-386.1
- 501 Cal. Civ. Code § 1708.86 (10), provides that "Nude" means visible genitals, pubic area, anus, or a female's postpubescent nipple or areola.
- 502 Cal. Civ. Code § 1708.86 (13) identifies a range of sexual activity that constitutes 'sexual conduct'
- 503 Cal. Civ. Code § 1708.86 (14)(c)(1)(B)
- 504 The legislation makes it clear that sexually explicit material is not newsworthy only because it depicts a public figure; Cal. Civ. Code § 1708.86 (14)(c)(2).
- 505 Cal. Civ. Code § 1708.86 (14)(d).
- 506 California SB- 564 (2019-2020). Text and legislative history available at *https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201920200SB564*. [Accessed 11 June 2021].
- 507 California SB- 564 (2019-2020), § 1708.86(g)(1)
- 508 NY Civ Rights L § 52-c
- 509 NY Consolidated Laws - Penal § 130
- 510 NY Civ Rights L § 50-f (1)(b).
- 511 The provisions define the medium
- 512 NY Civ Rights L § 50-f (2)(b).
- 513 NY Civ Rights L § 50-f (3).
- 514 NY Civ Rights L § 50-f (8).
- 515 NY Civ Rights L § 50-f (2)(b).
- 516 NY Civ Rights L § 50-f (2)(d)(ii).
- 517 NY Civ Rights L § 50-f (2)(d)(iii).
- 518 E.g. New York Assembly Bill 8155 which was introduced in 2018 and failed to pass, proposed a cause of action for unauthorised commercial use of Deepfake similar in nature and scope to that now found in section 50 of the New York Civil Rights Law.
- 519 AB 1280 – Crimes Deceptive Recordings. A copy of the Bill and legislative history are available at *https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201920200AB1280* [Accessed 11 June 2021] The Bill also proposed a third offence of non-consensual creation and dissemination of a deepfake within 60 days of an election with the intention of coercing or deceiving any voter into voting for or against a candidate or measure in that election.
- 520 Bill H.3366, 191st session (2019-2020). Introduced 22 January 2019. Text and legislative history available at *https://malegislature.gov/Bills/191/H3366*. [Accessed 12 June 2021]
- 521 S. 3805 – 115th Congress (2017-2018), introduced Dec 2018. Text and legislative history available at *https://www.congress.gov/bill/115th-congress/senate-bill/3805*. [Accessed 11 June 2021]
- 522 Bill H.1755 – 192nd Session (2021). Introduced 29 March 2021. Text and legislative history available at *https://malegislature.gov/Bills/192/H1755*. [Accessed 12

June 2021].

- 523 H.R. 3230 – 116th Congress (2019-2020). Text and legislative history available at *https://www.congress.gov/bill/116th-congress/house-bill/3230.* [Accessed 11 June 2021]
- 524 The definition of this is set out in Part I of this report.
- 525 A number of other Bills that proposed similar reporting requirements have been introduced but failed to pass, see: H.R. 3600 – 116th Congress (2019-2020). Text and legislative history available at h*ttps://www.congress.gov/bill/116th-congress/house-bill/3600; HR 3494* – 116th Congress (2019-2020). Text and legislative history available at https://www.congress.gov/bill/116th-congress/house-bill/3494. [Accessed 14 Jun 2021]
- 526 It should be noted that legislation has since been enacted that imposes similar notification and reporting requirements on the Director of National Intelligence; see 50 USC c 45 § 3369a (2021)
- 527 Hoofdzin van artikel 150 Wetboek van Burgerlijke Rechtsvordering (hierna: Rv).
- 528 Ahsmann, M.J.A.M. (2020), D*e weg naar het civiele vonnis.* Den Haag: Boom juridisch 2020, p. 312-313.
- 529 Ahsmann, M.J.A.M. (2020), *De weg naar het civiele vonnis.* Den Haag: Boom juridisch 2020, p. 85-86 en 276-278.
- 530 Artikel 21 Wetboek van Burgerlijke Rechtsvordering.
- 531 https://www.theverge.com/2021/3/11/22323271/wombo-ai-memes-deepfake-app-lip-sync
- 532 https://www.deepnude.to
- 533 See further: Bailey, J., Burkell, J., Gosse, C., & Steeves, V. (2020). AI and Technology-Facilitated Violence and Abuse. *Artificial Intelligence and the Law in Canada (Toronto: LexisNexis Canada, 2021).*
- 534 https://nypost.com/2020/09/13/phone-cameras-are-color-correcting-the-west-coast-wildfires/
- 535 https://www.youtube.com/watch?v=TWFaErx_nMs
- 536 https://www.youtube.com/watch?v=CF_eokMCW20
- 537 https://nos.nl/artikel/2373770-om-zet-voor-het-eerst-online-lokprofiel-in-verdachte-jeugdprostitutie-opgepakt
- 538 https://www.youtube.com/watch?v=2du6dVL3Nuc
- 539 https://www.nytimes.com/2020/08/03/technology/fawkes-tool-protects-photos-from-facial-recognition.html
- 540 https://www.europol.europa.eu/publications-documents/malicious-uses-and-abuses-of-artificial-intelligence
- 541 https://contentauthenticity.org/
- 542 Situatie mei 2021
- 543 See, e.g. https://www.tandfonline.com/doi/full/10.1080/00913367.2020.1810595.
- 544 https://osf.io/cdfh3/
- 545 https://journals.sagepub.com/doi/pdf/10.1177/1940161220944364
- 546 Kwok, A. O. J., & Koh, S. G. M. (2020). Deepfake: A social construction of technology perspective. *Current Issues in Tourism,* 24:13, 1798-1802, DOI: 10.1080/13683500.2020.1738357
- 547 Maddocks, S. (2021). Feminism, activism and non-consensual pornography: analyzing efforts to end "revenge porn" in the United States. *Feminist Media Studies*, 1-16.
- 548 https://journals.sagepub.com/doi/full/10.1177/0964663920947791
- 549 https://datasociety.net/library/deepfakes-and-cheapfakes/
- 550 Keats Citron, D. (2018). Sexual privacy. *Yale LJ*, 128, 1870.
- 551 Perot, E., & Mostert, F. (2020). Fake it till you make it: an examination of the US and English approaches to persona protection as applied to deepfakes on social media. *Journal of Intellectual Property Law & Practice*, 15(1), 32-39.
- 552 Westerlund, M. (2019). The emergence of deepfake technology: A review. T*echnology Innovation Management Review,* 9(11).

Deepfakes: De juridische uitdagingen van een synthetische samenleving

Design by Julia Janssen

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