**TIEN VRAGEN OVER DEEPFAKES**

**Wat zijn deepfakes?**

Deepfakes zijn beelden, geluid of tekst gemaakt door AI of door mensen met behulp van AI. Het is dus nep. Het is vaak heel lastig om te beoordelen of een deepfake echt is of niet.

Tekst die gegeneerd wordt door ChatGPT, Copilot, Perplexity is dus ook deepfake. In de in de volksmond wordt meestal verwezen naar afbeeldingen of video's, zoals: de virale imitatie van acteur Tom Cruise op TikTok; Buzzfeed's Barack Obama deepfake, waarin de voormalige president zich lijkt te begeven in onkarakteristieke godslastering; deepfake video's van Poetin en Zelensky die zijn verschenen als propaganda tijdens de oorlog in Oekraïne, nepporno in Zuid-Korea.

Een ander ernstig voorbeeld is de financieel medewerker die dacht dat hij een videocall had met diverse medewerkers van zijn kantoor en toen $ 25 miljoen overmaakte aan cybercriminelen. https://edition.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html

Het is een groeiend probleem, wat veel slachtoffers eist.

**Zijn deepfakes rechtmatig?**

Dat kan zeker. Het hangt af van de omstandigheden, toestemming en context. In de reclame wordt al jaren veelvuldig gebruik gemaakt van deepfakes, zodat topsporters niet meer hoeven te acteren. Dit soort gebruik is echter met toestemming van de geportretteerde gemaakt.

Ook in Hollywood films wordt steeds meer gebruik gemaakt van door AI gemaakte beelden of geluid.

Andere voorbeelden zijn dat historische foto’s, of kunstwerken met behulp van AI worden bewerkt tot bewegende beelden.

Vermeldenswaardig zijn ook deep therapy toepassingen of apps zoals FaceApp, of bepaalde filters in Snapchat of Instagram.

**Zijn deepfakes strafbaar?**

Er kan op allerlei manieren een strafbaar feit gepleegd worden met deepfakes, of onrechtmatig worden gehandeld. Denk aan oplichting, fraude, phishing, sextortion, misleiding, inbreuk op intellectuele eigendomsrechten en natuurlijk schending van privacy. 95% van de deepfakes is non consensual porn. Dat is strafbaar en onrechtmatig, zodra je dat verspreidt.

Er wordt echter ook al jaren gelobbyd om de strafbaarheid van deepfakes te vergroten om zo de positie van slachtoffers te versterken. Zo is het aanbieden van nudify apps nog niet strafbaar in Nederland. Nudify apps zijn apps waarmee foto’s van mensen (voornamelijk vrouwen) zo bewerkt worden dat het resultaat een deepfake is waarop ze ontkleed te zien zijn met bijzonder veel detail. Het gebruik van die deepfakes kan echter wel strafbaar en onrechtmatig zijn.

**Hoe zit het met post-mortem deepfakes?**

In Nederland is daar geen specifieke wet voor. De nabestaanden zullen zich kunnen beroepen op het portretrecht (artikel 19-21 Auteurswet) van de overledene, als diens portret is gebruikt. Dan komt het aan op een afweging van het redelijk belang.

Bij voice clones zal een beroep op AVG niet mogelijk zijn, als het de stem van een overledene betreft. Dan biedt de onrechtmatige daad mogelijk een grondslag voor een verbod en schadeclaim.

Een voice clone maken van iemands stem is op zichzelf geen strafbaar feit. Wanneer deze opname als instrument wordt gebruikt (of wordt gepoogd) om anderen geld afhandig te maken, dan weer wel.

In Amerika komen er steeds meer (federale) wetten die dit regelen. Voor een overzichtsartikel uit 2023 naar de situatie in andere landen verwijs ik u naar https://ipkitten.blogspot.com/2023/09/guest-post-deepfake-it-till-you-make-it.html.

**Welke regels zijn er in de AI Act over deepfakes?**

In de AI Act staat dat providers van AI systemen die deepfakes kunnen maken, deze moeten markeren, zodat dit gedetecteerd kan worden.

Art. 50 (2): Aanbieders van AI-systemen, met inbegrip van AI-systemen voor algemene doeleinden die synthetische audio-, beeld-, video- of tekstinhoud genereren, zorgen ervoor dat de output van het AI-systeem in een machineleesbaar formaat wordt gemarkeerd en als kunstmatig gegenereerd of gemanipuleerd kan worden gedetecteerd. Aanbieders zorgen ervoor dat hun technische oplossingen doeltreffend, interoperabel, robuust en betrouwbaar zijn voor zover dit technisch haalbaar is, rekening houdend met de specifieke kenmerken en beperkingen van de verschillende soorten inhoud, de kosten van tenuitvoerlegging en de algemeen erkende stand van de techniek, zoals die eventueel in relevante technische normen tot uiting komt. Deze verplichting is niet van toepassing voor zover de AI-systemen een ondersteunende functie voor standaardbewerking uitvoeren of de door de gebruiker verstrekte invoergegevens of de semantiek daarvan niet wezenlijk wijzigen, of wanneer zij wettelijk zijn toegestaan om strafbare feiten op te sporen, te voorkomen, te onderzoeken of te vervolgen.

Overweging 133 van de AI-Act stelt dat “dergelijke technieken en methoden voldoende betrouwbaar, interoperabel, doeltreffend en robuust moeten zijn, voor zover dit technisch haalbaar is, rekening houdend met beschikbare technieken of een combinatie van dergelijke technieken, zoals watermerken, metagegevensidentificaties, cryptografische methoden om de herkomst en authenticiteit van inhoud aan te tonen, logmethoden, vingerafdrukken of andere technieken, voor zover van toepassing”.

Opvallend is dat deze verplichting alleen voor providers geldt. Ook is het verwijderen van het watermerk niet strafbaar gemaakt. Dat is een gemiste kans.

**Wanneer geldt die verplichting?**

Vanaf 2 augustus 2026

**Wat is het risico?**

Op niet-naleving staat een boete van maximaal 15 miljoen euro (artikel 99, lid 4, onder g).

**Hoe moeten providers dit doen?**

Op dit moment zijn er verschillende vormen ontwikkeld waarmee synthetische media gemarkeerd kunnen worden. Zoals watermerken en metadata, cryptografische methoden, van de C2PA standaard https://c2pa.org/.

**Hoe zit het met bewijs in de rechtszaal?**

Rechters, het OM en advocaten moeten zich goed realiseren dat het tegenwoordig heel gemakkelijk is om bewijs te vervalsen.

In 2019 heeft een advocaat die optrad in een voogdijgeschil in het Verenigd Koninkrijk voor de vader die in Dubai woonde, met succes audiobewijs aangevochten waarin de vader als gewelddadig en agressief werd afgeschilderd. Door toegang te krijgen tot de audiobestanden konden forensische experts aantonen dat de opname een 'deepfake' was, die de moeder had samengesteld met behulp van online hulpfora. Audiobewijs is nog steeds relatief snel en gemakkelijk overtuigend te vervalsen in vergelijking met video en foto's en er zijn al hoogwaardige apps op de massamarkt vrij verkrijgbaar om 'stemklonen' te maken.

Onderzoekers van de Universiteit van Tilburg stellen in een rapport uit 2021 voor om zorgplichten op te leggen aan advocaten, politie, aanklager en zelfs rechters om al het ingediende bewijs te laten verifiëren door een onafhankelijke forensische expert. Deze opties zijn echter kostbaar en tijdrovend, vooral gezien het tekort aan relevante deskundigen. Een nationaal expertisecentrum is nog toekomstmuziek.
Het is wachten op rechters die met sancties komen voor het overleggen van vals bewijsmateriaal.

**Wat kan je doen als je slachtoffer bent geworden van deepfake?**

Een aantal stappen zijn belangrijk.

1. Leg het bewijs van de deepfake vast, door screenprints te maken;
2. Rapporteer het bij de provider via wiens dienst het wordt gedeeld (Whatsapp heeft bijvoorbeeld een rapporteerfunctie);
3. Praat er over met iemand die je vertrouwt;
4. Dien een handhavingsverzoek in bij de Autoriteit Persoonsgegevens of doe aangifte bij de politie.
5. Andere organisaties die je kunnen helpen zijn Slachtofferhulp en Helpwanted. Bij Slachtofferhulp kun je terecht voor emotionele hulp, maar ook voor rechtshulp of praktische informatie over aangifte doen en het strafproces. Helpwanted is er speciaal voor advies bij (online) seksueel misbruik. Dit kan anoniem.
6. Neem contact op met een advocaat.