WAT ZIJN DEEPFAKES?

Wat is echt en niet echt? Door **deepfakes** wordt de lijn tussen realiteit en fantasie dunner. Maar wat zijn deepfakes juist?

Kom er meer over te weten in deze brAInfood. We focussen op de **werking** ervan, tonen enkele **voorbeelden** en duiden op de **risico's** en **voordelen**.

Kenniscentrum Data & Maatschappij (2021). Wat zijn deepfakes? brAInfood van het Kenniscentrum Data & Maatschappij. Brussel: Kenniscentrum Data & Maatschappij.

Deze brAInfood is beschikbaar onder een CC by 4.0 licentie.

WAT ZIJN DEEPFAKES?

'Deepfakes' zijn **beelden, geluiden en teksten** die door **artificiële intelligente** worden **gecreëerd** of **gemanipuleerd**. Hiervoor maakt het gebruik van een reeks bestaande beelden, audio of teksten, waarvan de software zichzelf bepaalde zaken aanleert (bv. hoe een gezicht eruitziet). Met deze informatie kan de software bijvoorbeeld een nieuwe afbeelding van een gezicht creëren.

Een **General adversarial network** (GAN) is een machine learning model waarin **twee neurale netwerken** met elkaar **concurreren** om **nauwkeuriger voorspellingen** te doen. Het ene netwerk genereert een syntetisch voorbeeld dat gelijkaardig is aan de reële data terwijl het andere probeert te onderscheiden of een voorbeeld reëel of syntetisch is. Op die manier verbeteren de twee netwerken zich voortdurend en zorgt de software voor nieuwe content die bijna niet te onderscheiden is van de originele content.

WAT ZIJN DE RISICO'S?

Deepfakes zijn vaak zo realistisch dat ze moeilijk van echte beelden te onderscheiden zijn. Ze kunnen de waarheid en realiteit vervormen.

Dit brengt risico's met zich mee als ze voor verkeerde doeleinden worden gebruikt. Deepfakes kunnen bijvoorbeeld:

- **nepnieuws** creëren en geloofwaardiger maken;
- politieke **campagnes verstoren;**
- **extremisme** aanwakkeren;
- bewijsstukken **vervalsen**;
- algemeen **wantrouwen** in bepaalde personen of media vergroten.

We zijn gewend dat (video)beelden de waarheid tonen en dankzij sociale netwerken is het gemakkelijk om (video)beelden massaal te delen. Dit maakt dat de valse informatie in deepfakes **grote groepen** mensen kan **beïnvloeden**. In de verkeerde handen heeft de technologie mogelijks grote gevolgen voor de democratie, veiligheid en internationale zaken.

WAT IS DE MEERWAARDE?

Deepfakes en GANs kunnen bijdragen aan de **entertainment-, kunst- en cultuursector, het onderwijs en de gezondheidszorg**: van het nabootsen van reeds lang overleden kunstenaars of artiesten tot het bewerken van video's zonder dat er nieuwe opnames nodig zijn.

Het AI-lab van Samsung in Moskou kon de Mona Lisa van Da Vinci omzetten in een video waarin de vrouw haar ogen, hoofd en mond beweegt. Deepfaketechnologie kan geschiedenis tot leven brengen, en onderwijs, kunst en cultuur **toegankelijker en interactiever** maken voor een breed publiek.

Met StyleGAN, een specifieke toepassing van GAN, kan je een beeld genereren in een bepaalde visuele stijl die je de technologie aanleert. Ze worden ingezet in **fashion design, web art, animaties, muziek, ...**

In de **gezondheidszorg** kunnen deepfakes realistische data reconstrueren, die onderzoekers helpen **nieuwe behandelingen** te ontwikkelen zonder echte patiëntendata te gebruiken. Een veelbelovend experiment is de ontwikkeling van kunstmatige MRI-scans van hersenen die helpen een algoritme te trainen op het herkennen van hersentumoren.

HOE ZE TE HERKENNEN?

Het herkennen van deepfake video's is niet altijd makkelijk, zeker als deze gemaakt zijn met professionele tools en budgetten.

Bij het bekijken ervan kan je het gevoel hebben dat er **iets 'mis'** is met de video die je ziet. Een algoritme kan namelijk moeilijk **lichaamstaal** en **subtiele menselijke expressies** genereren en in de juiste context plaatsen. Let daarom op onnatuurlijke gezichtsexpressies en oogbewegingen (bv. te veel of te weinig knipperen met de ogen), verkeerde perspectieven (van bv. een neus), etc. bij het kijken van een video.

Er zijn ook **(open-source) deepfake opsporingstools** die gebruik maken van verschillende technieken om deepfake video's te ontmaskeren. Zo heb je bijvoorbeeld een algoritme dat digitale hashes en metadata van een bestand kan analyseren op echtheid en originaliteit.

Dergelijke tools werken naar behoren, maar dankzij de vooruitgang in de technologie, zal het maken van deepfakes en het opsporen ervan nog lange tijd een **kat-en-muis spel** zijn.