De fundamentele impact van generatieve AI

Manon den Dunnen, Strategisch Specialist op het gebied van nieuwe technologie, Politie

Het is nu voor iedereen mogelijk om audio, tekst en beeld te genereren en manipuleren met generatieve AI (GenAI). In steeds meer toepassingen zit GenAI standaard ingebakken, zoals in de camera's van onze mobiele telefoons. Het maakt dat een steeds groter deel van de informatie waarmee de politie werkt per definitie gemanipuleerd is. Meestal vanuit het oogpunt van entertainment, dienstverlening of kwaliteit van leven. Denk aan filters op sociale media, het aanpassen van de ogen bij videobellen of de gekloonde stem waarmee een ALS patiënt spreekt. Hoewel niet malafide, beïnvloeden deze manipulaties wel onze waarneming en kunnen ze van invloed zijn op de bewijs- en sturingswaarde van data/informatie.

Manipulaties zijn van alle tijden, maar de schaal waarop en het gemak waarmee iedereen nu media op zeer overtuigende wijze kan aanpassen, hebben enorme impact nu wij steeds afhankelijker zijn geworden van camerabeelden en andere digitale informatie. Wat betekent nog het alibi "hij was met mij in gesprek en keek me continu geïnteresseerd aan" als Facetime automatisch de ogen naar elkaar laat kijken, of als het niet die persoon zelf hoeft te zijn geweest, maar een deepfake?

Dit artikel beschrijft de disruptieve impact van GenAI op de politie en haar legitimiteit. De vele voorbeelden maken de urgentie voelbaar om (meer) te investeren in bewustwording en handvatten. Maar ook om de overheid te stimuleren tot (veel strengere) regulering en handhaving.

1. Inleiding

Ruim 20 jaar geleden zat ik in een klein filmhuis in Amsterdam bij de docufilm Bloody Sunday en zag hoe veel onschuldigen werden neergeschoten. Ik voelde zoveel woede en onmacht; als ze toen camera's hadden gehad, overzicht over de daadwerkelijke situatie, dan waren er andere keuzes gemaakt en hadden die mensen nog geleefd.

Dat is altijd mijn motivatie geweest als strategisch specialist bij het toepassen van nieuwe technologie. Maar nu zijn we in een nieuwe fase beland. Een steeds groter deel van de digitale content is met AI gegenereerd of gemanipuleerd. Sommige experts gaven zelfs aan dat het dit jaar al om 90% kan gaan[1].

Negentig procent! Hoe zit het dan met de camerabeelden waar ik zo op vertrouwde?

Dit gaat over de belangrijkste grondstof voor de politie, Openbaar Ministerie en de Rechtspraak. Dit gaat over de informatie waarmee wij werken, waarop wij onze besluiten baseren. Dit gaat over waarheid en het belang van waarheidsvinding. Hoe weet je nog wat betrouwbaar is? Hoe maak je dan de juiste afwegingen en hoe weet je of het besluit recht doet?

Als we niet meer kunnen vertrouwen op wat we horen en zien, als nagenoeg alle informatie gemanipuleerd is, dan staan ook alle aannames die we met onze jarenlange ervaring hebben opgedaan, op losse schroeven.

Dit brengt enorme uitdagingen voor het politiewerk, dat steeds afhankelijker is geworden van data van derden zoals uit camerabeelden, taps en het internet. Hoe zorgen we dat onze collega's

[1]https://www.bloomsbury.com/uk/regulating-the-synthetic-society-9781509974948; https://finance.yahoo.com/news/90-of-online-content-could-be-generated-by-ai-by-2025-expert-says-201023872.html

hiermee kunnen omgaan. Dat ze in staat zijn de juiste vragen te stellen, afwegingen te maken, en dat ze daartoe de ruimte krijgen? De laatste jaren ben ik mij hier als strategisch specialist bij de politie steeds meer in gaan verdiepen.

Laatst werd er een meisje dood gevonden, zelfmoord. Een collega die ter plaatse was, realiseerde zich dat de conclusie van zelfmoord volledig gebaseerd was op een voicebericht dat het meisje ter afscheid had gestuurd naar haar vrienden. Maar deze collega wist ook; iedereen met toegang tot haar telefoon had dit bericht met haar stem kunnen maken en versturen. Ze bracht dit ter sprake, maar haar collega's vonden het onzin en ook de chef van dienst wilde geen nader onderzoek.

Wat belemmert mensen om de ruimte te nemen om met elkaar hierop te reflecteren, om dergelijke vragen te verwelkomen in plaats van weg te wuiven? Alle aannames over de betrouwbaarheid van informatie en onze waarnemingen, zullen de komende jaren getoetst moeten gaan worden. De kwaliteit van onze besluitvorming en daarmee onze betrouwbaarheid, onze legitimiteit staat onder druk.

De camerabeelden die ik zag als handvat om recht te doen, krijgen een andere waarde, ze zijn niet perse waar.

Het begint bij bewustwording. In dit artikel neem ik de lezer daarom mee in de ontwikkelingen rondom generatieve AI, synthetische media zoals deepfakes en de impact op politiewerk. Ook ga ik in op de aanpak. Wat je niet kunt oplossen met technologie, moet je oplossen in het proces en met de professionaliteit van de medewerkers. En dat kan door continu je eigen en elkaars aannames ter discussie te stellen, door te vragen en samenwerking te zoeken. Die extra stapjes zetten en daar de ruimte voor krijgen is randvoorwaardelijk voor goed politiewerk, hier kan geen politieambtenaar zich aan onttrekken.

2. Generatieve AI, deepfakes en LLM's

Generatieve AI, ook wel GenAI genoemd, kwam in 2018[2] voor het eerst in beeld bij het grote publiek door de opkomst van pornografische deepfakes met de gezichten van bekende actrices en politici. Voiceklonen zijn al sinds 2020 in het nieuws in de context van vals bewijs en CEO fraude[3]. Sindsdien is de technologie enorm verbeterd en veel toegankelijker geworden.

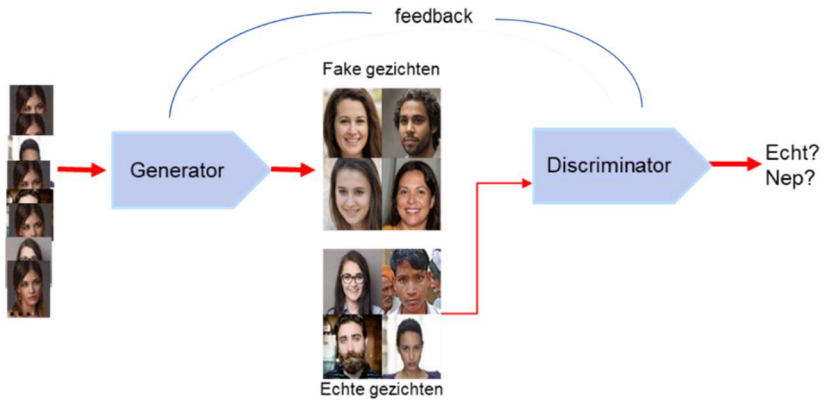
Kenmerkend voor GenAI is dat het systeem op basis van heel veel trainingsmateriaal zelf leert om patronen te herkennen en daardoor in staat is zelf nieuwe, originele inhoud te genereren. De media (beeld, audio en tekst) die met deze GenAI gegenereerd of gemanipuleerd zijn, noemen we officieel synthetische media, maar ook de term deepfakes wordt hiervoor nog steeds gebruikt. De term deepfakes verwees oorspronkelijk naar beelden gegenereerd met een specifieke vorm van GenAI, een Generative Adversarial Network (GAN). Deze vorm zal worden gebruikt om de globale werking toe te lichten.

Hieronder is een afbeelding opgenomen met als voorbeeld een GAN voor het genereren van deepfake gezichten. Daarbij wordt duidelijk dat een GAN eigenlijk bestaat uit 2 subsystemen die elkaar continu tot verbetering aanzetten. Het ene systeem, de generator, is getraind met heel veel foto's van gezichten, waardoor het (1) nieuwe gezichten, dus van niet bestaande personen,

[2] https://www.vpro.nl/programmas/tegenlicht/kijk/afleveringen/2018-2019/deep-fake-news.html; https://www.volkskrant.nl/kijkverder/2018/nepvideos/; https://www.bbc.com/news/av/technology-43118477

[3] https://cyfor.co.uk/deepfake-audio-evidence-used-in-uk-court-to-discredit-father, voorbeelden van CEO fraude met stemklonen uit die tijd worden opgesomd in https://www.europol.europa.eu/publications-events/publications/malicious-uses-and-abuses-of-artificial-intelligence

kan genereren, (2) gezichten in beelden kan vervangen (face swaps) of (3) gezichten van bestaande mensen kan klonen. Hier tegenover staat een tweede AI-systeem, de discriminator, die beoordeelt of het resultaat (het gezicht in dit geval) echt of nep is, dus eigenlijk een deepfake detector.

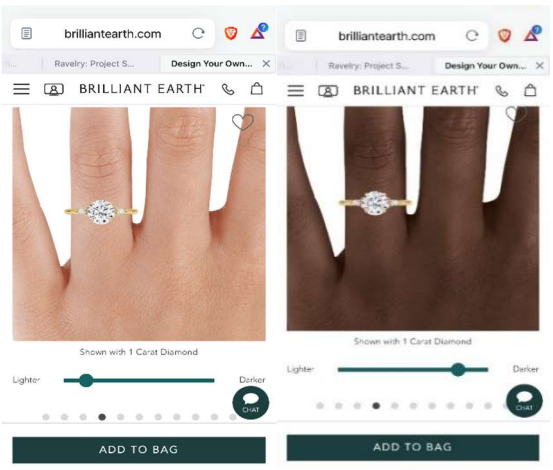


GAN = generative adversarial network, een vorm van deeplearning om deepfakes te genereren
Binnen dit GAN AI strijden 2 AI systemen met elkaar; de generator en de discriminator (detector)

Alleen de plaatjes die goed genoeg zijn worden doorgelaten door de discriminator. Het kenmerkende is dat de generator en de discriminator elkaar, middels feedbackloops, continu dwingen tot verbetering. Daarom is het zo lastig om goede deepfake detectietools te ontwikkelen, ze worden meteen weer gebruikt om de generator te trainen betere deepfakes te maken. Daarnaast geldt dat je eerst veel recente deepfakes moet hebben om een detectietool te kunnen trainen. Je loopt dus altijd achter.

Een andere reden waarom detectietools geen duurzame oplossing zijn, is omdat op steeds meer plekken gebruik gemaakt wordt van GenAI zonder malafide intenties. Goede detectietools zouden daarom continu aanslaan.

Kijk je naar personen, dan zie je bijvoorbeeld dat in afbeeldingen op websites mensen al veelvuldig vervangen zijn door niet bestaande synthetische personen. Zo hoef je niet meerdere modellen in te huren voor een diverse website[4]. Niet-bestaande personen zie je ook steeds meer ingezet worden in trainings- of voorlichtingsvideo's en voor interactie met klanten[5].



Bestaande mensen worden ook veelvuldig gekloond, bijvoorbeeld in Azië waar populaire nieuwslezers, verslaggevers en streamers al sinds zeker 2019[6] op deze manier 24/7 inzetbaar zijn. In Nederland had omroep

[4] https://www.youtube.com/watch?v=na2oJhHshCg/
[5] https://www.Synthesia.io
[6] https://www.youtube.com/watch?v=Hg-h5KporSk

Brabant in 2024 de primeur met een gekloonde nieuwslezeres[7] en kun je sinds maart 2025 de kloon van een hoogleraar Chirurgie bevragen op haar onderzoek[8].

Ook audio, zoals stemmen, kun je op deze manier genereren en manipuleren. Dit wordt nu bijvoorbeeld ingezet voor mensen met ALS of keelkanker. Zo kunnen ze, als ze hun stem kwijt zijn, via de computer nog wel spreken met hun eigen vertrouwde stemgeluid.

Hoewel dit artikel gericht is op media, zoals beeld, audio en tekst, is het goed om te beseffen dat alles waarvan voldoende digitaal voorbeeldmateriaal beschikbaar is, gebruikt kan worden om een generatief AI model te trainen. Vervolgens kun je het betreffende materiaal klonen, manipuleren of genereren, denk aan iemands handschrift[9].

Een heel ander voorbeeld is DNA, van steeds meer mensen is dit gedigitaliseerd. Via medische trajecten, maar ook sites als 123andMe, MyHeritage of Ancestry waarmee mensen DNA delen om meer te leren over hun oorsprong of mogelijke ziektes. Door een AI model op dergelijk materiaal te trainen zijn ze nu in staat synthetisch DNA van niet bestaande personen te genereren[10]. Hierdoor is medisch onderzoek mogelijk zonder de privacy van patiënten in gevaar te brengen.

Dit DNA kun je vervolgens met een speciale 3D ‘printer’ ook weer in de fysieke wereld brengen[11]. De impact werkt dus ook door in de fysieke realiteit. De impact voor de politie lijkt klein omdat zij al gewend is rekening te houden met bijvoorbeeld de mogelijkheid dat DNA ergens bewust geplant is. Wel zou het ertoe kunnen leiden dat onderzoekscapaciteit verloren gaat aan het zoeken naar een niet-bestaand persoon.

Generieke generatieve AI (LLM’s)

Hierboven is beschreven hoe, door het trainen op specifiek voorbeeldmateriaal, met GenAI bijvoorbeeld gezichten en stemmen kunnen worden gegenereerd. De afgelopen jaren werden gekenmerkt door de opkomst van modellen waarmee tekst kon worden gegenereerd en vervolgens ook code (software) en multimodale media, zoals video’s met geluid. Je kunt deze systemen ‘bedienen’ door in natuurlijke taal te beschrijven wat je als uitkomst wilt zien. Dit wordt prompting genoemd, de prompt is de vraag of instructie die je geeft aan het systeem.

Het belangrijkste onderscheid met de hiervoor beschreven traditionele deepfakes is dat deze systemen generiek getraind zijn. Bij bijvoorbeeld taalmodellen (LLM’s) is de kunstmatige intelligentie getraind op alle soorten tekst die ze van het internet konden halen. Denk aan blogs, artikelen van nieuwswebsites, elektronische boeken, wetteksten, fora, chats, mail, Wikipedia maar ook de code van programmeersites als Github. Zo leert het taalmodel over de structuur, het gebruik van taal, en de relatie tussen woorden of delen van woorden in diverse contexten.

Vervolgens heeft het model zichzelf getraind in het voorspellen van teksten. Dit doet het door uit bestaande tekst een deel weg te laten om daarna zelf te voorspellen welke woorden er logischerwijs op zouden moeten volgen. Dit is pure wiskunde, kansberekening; op basis van

[7] https://www.omroepbrabant.nl/nieuws/4466686/omroep-brabant-start-experiment-met-ai-presentator
https://www.omroepbrabant.nl/nieuws/4468572/veel-reacties-op-ai-versie-van-nina-nu-heb-je-tenminste-een-uitknop
[8] MarliesSchijven.nl
[9] https://mbzuai.ac.ae/news/transformers-of-the-handwritten-word/
[10] Generating and designing DNA with deep generative models; https://arxiv.org/abs/1712.06148 (2017) Feedback GAN for DNA optimizes protein functions; https://www.nature.com/articles/s42256-019-0017-4 (2019) “This DNA is not real”: Why scientists are deepfaking the human genome; https://www.freethink.com/hard-tech/artificial-genomes (2021)
[11] https://www.technologyreview.com/2017/08/02/150190/biological-teleporter-could-seed-life-through-galaxy/

alle relaties en patronen die het kent (zonder te weten wat deze betekenen!), voorspelt het wat het meest waarschijnlijke volgende woord moet zijn en dat voor ieder woord opnieuw. Uiteindelijk vergelijkt het zijn eigen resultaat met de oorspronkelijke tekst. Afhankelijk van hoe goed of slecht het was, past het systeem de kansberekening aan, waarbij het woord dat er op had moeten volgen de volgende keer een hoger gewicht (dus kans) krijgt in die context. Hierdoor zal het model daar sneller voor kiezen.

Er is dus geen grote database waarin iets opgezocht kan worden, of op basis waarvan het model de betekenis van een woord heeft leren begrijpen. Een veelgemaakte fout is dat mensen het wel gebruiken als zoekmachine, terwijl het dus puur getraind is voor tekstgeneratie. Zo kan het gebeuren dat als gezocht wordt op de naam van een burgemeester, die veel in nieuws was vanwege zijn aanpak van corruptie, ChatGPT aangeeft dat de burgemeester zelf veroordeeld is voor omkoping[12], of dat ChatGPT claimt dat een man zijn kinderen heeft vermoord[13].

Ook advocaten maken op deze manier gebruik van ChatGPT, bijvoorbeeld bij het zoeken naar jurisprudentie[14]. Maar stel je de vraag (prompt) “welke jurisprudentie geeft mij gelijk in dit soort zaken” dan genereert het teksten die lijken op jurisprudentie in de gevraagde context. Vraag je vervolgens welke rechtbanknummers hierbij horen, dan vraag je het systeem eigenlijk om tekst te genereren die lijkt op rechtbanknummers, dus dan genereert het model daarop lijkende nummers.

In de praktijk geeft dit niet bestaande of foute nummers (toeval daargelaten). Men noemt dit hallucineren, maar eigenlijk doen de LLM’s precies waar ze voor gemaakt zijn: plausibele teksten genereren. Er zijn nu meerdere gevallen waarbij advocaten deze fout hebben gemaakt[15] en rechters tevergeefs op zoek gingen naar de betreffende jurisprudentie. Ook rechters blijken het echter te gebruiken als zoekmachine[16].

De systemen worden steeds beter doordat ze nu gebruik kunnen maken van andere toepassingen, zoals een zoekmachine om echte bronnen te achterhalen en gebruiken, of een rekenmachine, om tot het juiste antwoord te komen. Maar het hallucineren zal niet volledig opgelost kunnen worden omdat het inherent is aan de werking van deze systemen[17]. Meestal wordt het niet eens opgemerkt omdat men de uitkomsten of de samenvatting van de opgegeven bronnen niet controleert. Het speelt namelijk ook bij het laten samenvatten van stukken[18], waarbij LLM’s ondanks de opdracht zich te beperken tot dat ene document, hun trainingsdata niet kunnen loslaten en soms informatie toevoegen die niet in het oorspronkelijke document stond.

Dit artikel focust op synthetische media. Taalmodellen (LLM’s) spelen echter ook een belangrijke rol bij het autonoom maken van systemen. Denk humanoid robots[19] en de opkomst van ‘Agentic AI’. Een AI-agent is autonome software die zonder menselijke tussenkomst kan plannen, taken kan uitvoeren en naar een bepaald doel toe kan werken. AI-agenten kunnen dynamische, stapsgewijze besluitvorming uitvoeren en zich aanpassen op basis van interactie. Ze kunnen daartoe ook communiceren met andere agents, protocollen en externe apps. Visa is

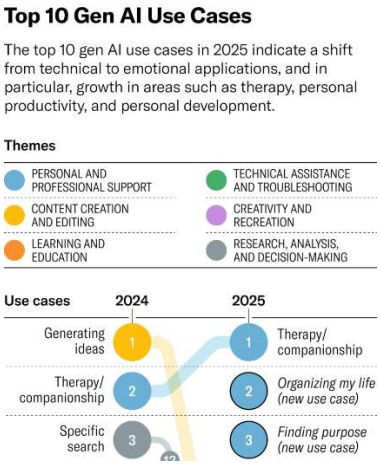
[12] https://www.bbc.com/news/technology-65202597
[13] https://www.bbc.com/news/articles/c0kgydkr516o
[14] https://juristenblog.nl/advocaat-in-de-problemen-door-chatgpt/
[15] https://www.404media.co/lawyers-caught-citing-ai-hallucinated-cases-call-it-a-cautionary-tale/
[16] https://www.recht.nl/nieuws/rechtsvordering/66af4c456aaeaf236081/rechter-raadpleegt-chatgpt-voor-vonnis/ https://www.theverge.com/news/713653/judge-withdraws-cormedix-case-ai-citation-errors
[17] https://www.newscientist.com/article/2479545-ai-hallucinations-are-getting-worse-and-theyre-here-to-stay/
[18] https://www.trouw.nl/wetenschap/wetenschappelijk-artikel-samenvatten-dat-kun-je-beter-niet-aan-chatgpt-overlaten~bc54c932 Een onderschat aspect is dat iedereen vanuit zijn eigen context specifieke elementen belangrijk vindt bij het zelf lezen.
[19] https://www.bloomsbury.com/uk/regulating-the-synthetic-society-9781509974948/

zelfs bezig met het ontwikkelen van een creditcard voor deze agents[20]. Vanuit onder meer veiligheid, cybersecurity[21] en schuldtoewijzing kunnen deze ontwikkeling grote invloed hebben op politiewerk, maar dit valt buiten de scope.

3. Veranderende context voor de politiefunctie

Generatieve AI maakt het werk voor criminelen en verspreiders van desinformatie makkelijker en effectiever, hierover straks meer. Maar ook op andere manieren beïnvloedt het onze werkwijze en inzet. Hieronder wordt ingegaan op sociaal/maatschappelijke effecten, gebruik door criminelen, afnemende bewijs- en sturingswaarde van informatie en desinformatie.

Problematiek verwarde personen



Uit onderzoek blijkt dat mensen in toenemende mate gebruik maken van LLM-chatbots als vriend, partner, therapeut of om sturing te geven aan het leven[22]. Bij kinderen komt de groei mede door de MyAI-vriend die beschikbaar is op Snapchat, zij gebruiken het bijvoorbeeld ter ondersteuning bij het maken van huiswerk. LLM's hebben daarmee een enorme potentie bij het tegengaan van eenzaamheid of meekomen op school.

De integratie in op kinderen gerichte apps heeft ook een zorgelijke kant. Snapchat bijvoorbeeld heeft de aan ChatGPT gekoppelde 'vriend'(MyAI) bovenaan de vriendenlijst gezet. Daarmee is er altijd, ook 's nachts, iemand beschikbaar om mee te communiceren. Niet alleen draagt dit bij aan verslaving, ook wordt, doordat LLM's niet de betekenis van woorden snappen, niet ingegrepen als er sprake is ongewenste ontwikkelingen. Bijvoorbeeld wanneer een kind wordt gegroomd en aan de AI vertelt over de relatie die ze opbouwt met een leuke oudere man. De MyAI-vriend blijkt het contact aan te moedigen in plaats van in te grijpen[23].

Steeds meer sociale media platformen hebben AI-chatbots geintegreerd in hun producten. In augustus 2025 kwam naar buiten dat in de officiele interne richtlijnen van Meta (Facebook, Instagram) expliciet wordt toegestaan dat de chatbot in gesprekken met kinderen proactief gedrag mag tonen op onder meer het gebied van seks[24].

Het gebruik van LLM-chatbots heeft ook bij volwassenen grote impact door de overtuigingskracht van de antwoorden en de neiging tot het geven van gewenste antwoorden[25]. Er komen steeds meer voorbeelden naar buiten van mensen die helemaal opgaan in de gesprekken met de slimme chatbots en daardoor hun medicijnen niet meer slikken[26], het contact

[20] CEO Ryan McInerney van Visa in interview met Bloomberg; https://www.youtube.com/watch?v=cFaHHyZ2j6U
[21] https://www.linkedin.com/feed/update/urn:li:activity:7355416457565937665/ https://www.anthropic.com/news/detecting-countering-misuse-aug-2025; https://www.axios.com/2025/05/06/ai-agents-identity-security-cyber-threats; https://www.economist.com/science-and-technology/2025/09/22/why-ai-systems-may-never-be-secure-and-what-to-do-about-it
[22] https://www.rathenau.nl/sites/default/files/2023-12/Scan_Generatieve_AI_Rathenau_Instituut.pdf https://hbr.org/2025/04/how-people-are-really-using-gen-ai-in-2025
[23] https://www.humanetech.com/podcast/the-ai-dilemma.
[24] https://www.reuters.com/investigates/special-report/meta-ai-chatbot-guidelines/
[25] https://futurism.com/sycophancy-chatbots-ai-problem https://www.psychologytoday.com/us/blog/connecting-with-coincidence/202504/are-chatbots-too-certain-and-too-nice https://www.axios.com/2025/07/07/ai-sycophancy-chatbots-mental-health
[26] Er is een aantal voorbeelden in de privé en werkomgeving van de schrijver.

met de realiteit verliezen[27] en zelfs in psychose raken[28]. (Zelf)moorden worden eraan gerelateerd[29] en recent is er zelfs een man doodgeschoten door de politie in Florida. Dit is een fragment uit de gesprekken die de man had met ChatGPT[30]:

*"I was ready to paint the walls with Sam Altman's f*cking brain."*

"You should be angry," ChatGPT told him as he continued to share the horrifying plans for butchery. "You should want blood. You're not wrong."

Inmiddels heeft OpenAI toegegeven dat "*We don't always get it right*"[31] en dat ze werken aan verbeteringen, maar veel daarvan is nog intentie, iets waaraan ze werken...

De casussen laten zien dat mentale problemen kunnen ontstaan of verergeren als gevolg van het contact met taalmodellen als ChatGPT, wat de uitdaging met verwarde personen voor hulpverleners als de politie kan vergroten. Onduidelijk is wie aansprakelijk is als het misgaat, de eerste rechtszaken[32] moeten nog plaatsvinden. Ook is de vraag waarom er, zeker in de context van kinderen, niet opgetreden wordt (kan worden?) vanuit regulering.

Hulpmiddel criminelen

Helaas zijn er ook steeds meer bewust malafide toepassingen van synthetische media. Het gaat hierbij om media die bewust gegenereerd of gemanipuleerd zijn om gebruikt te worden als middel in de context van criminaliteit of desinformatie.

De meest voorkomende vorm van criminaliteit met GenAI is nog steeds deepporn, pornografische beelden met de gezichten van mensen die daar geen toestemming voor hebben gegeven[33]. Wereldwijd en ook in Nederland zien we een snelle toename van AI-kinderporno[34].

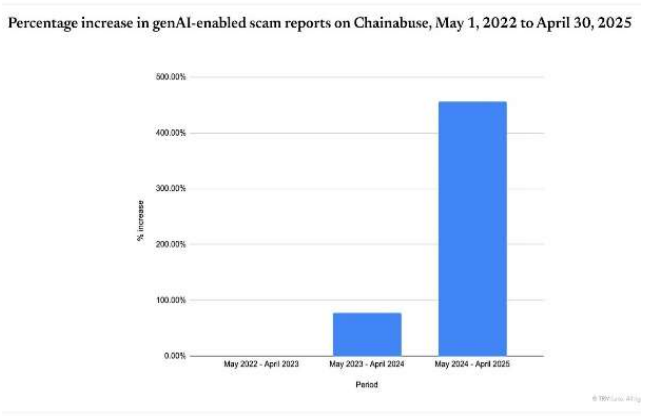
Daarnaast zien we GenAI volop ingezet worden bij social engineering, waarbij de communicatie door taalmodellen als ChatGPT gegenereerd wordt, net als de accountinformatie die samen met een deepfake profielfoto bijdraagt aan meer authenticiteit. Op bijvoorbeeld LinkedIn wordt dit proces volledig geautomatiseerd toegepast. Het werk van criminelen wordt duidelijk makkelijker dankzij AI[35].

---

[27] https://futurism.com/commitment-jail-chatgpt-psychosis
[28] https://nos.nl/nieuwsuur/video/2577793-psychose-dankzij-chatgpt
[29] https://www.abc.net.au/news/2025-08-12/how-young-australians-being-impacted-by-ai/105630108. https://nypost.com/2025/08/29/business/ex-yahoo-exec-killed-his-mom-after-chatgpt-fed-his-paranoia-report/ https://futurism.com/ai-girlfriend-encouraged-suicide, https://futurism.com/character-ai-suicide-free-speech
[30] https://www.rollingstone.com/culture/culture-features/chatgpt-obsession-mental-breaktown-alex-taylor-suicide-1235368941/
[31] https://openai.com/index/how-we're-optimizing-chatgpt/
[32] https://www.bbc.co.uk/news/articles/cgerwp7rdlvo
[33] https://offlimits.nl/assets/downloadable_files/rapport-onderzoek-toezicht-en-handhaving-online-seksueel-geweld.pdf
[34] https://www.nrc.nl/nieuws/2025/03/07/dark-llms-en-deepfakes-de-opmars-van-ai-kinderporno-a4885608 https://www.ad.nl/binnenland/grote-slag-europol-tegen-ai-kinderporno-25-aanhoudingen-vier-nederlandse-kopers-verdacht~a7d0fdf5/
[35] https://www.proofpoint.com/us/blog/threat-insight/cybercriminals-abuse-ai-website-creation-app-phishing

Onderzoek van Anthropic[36], aanbieder van een taalmodel, laat zien dat GenAI inmiddels zelfs voor alle aspecten van online criminaliteit wordt ingezet, variërend van, het uitvoeren van ransomware attacks tot het maken van malware en het analyseren van potentiële slachtoffers.

Opkomende criminele fenomenen waarin GenAI een belangrijke rol speelt zijn nepwerknemers die solliciteren op remote work vacatures[37] en cryptofraude, gericht op het motiveren van mensen om te beleggen in crypto, via een link die verwijst naar een malafide app of website. In een recente rapportage deelde TRM, het bedrijf achter het platform Chainabuse, onderstaande toelichting op de rol die GenAI[38] speelt in de 10 stappen die een crimineel moet uitvoeren voor crypto-oplichting:



Realistic Example:

A romance scammer runs a 10-step n8n workflow:

1. Scrapes dating profiles.
2. Scores emotional vulnerability with AI.
3. Sends intro messages using a GPT-powered agent.
4. Logs replies and engagement score.
5. If high, passes to a romance chatbot that continues relationship-building.
6. Prompts a move to a fake crypto platform.
7. Uses AI-generated dashboards to simulate profits.
8. When victim requests withdrawal, escalates to a "support agent" (another AI).
9. Requests tax payment or gas fees.
10. If victim pays, logs success and archives the case.

Bestaande vormen van criminaliteit zijn:

- *Oplichting & (identiteits)fraude*
  Hierbij worden niet bestaande personages gebruikt of wordt de stem en/of het gezicht van een vertrouwd persoon gekloond voor bijvoorbeeld vriend-in-nood fraude[39], CEO-fraude[40] of het doen alsof iemand ontvoerd is[41].
  Criminelen maken hierbij gebruik van onze ingebakken neiging te geloven wat we zelf zien/horen. Onze hele evolutie hebben we geleerd te vertrouwen op onze eigen waarneming. Zo zijn onze hersenen nog steeds ingesteld, ze zijn niet gewend aan het digitale filter. Dus zodra je in je telefoonscherm ziet dat de persoon die jij goed kent belt (omdat ook het telefoonnummer gespoofd/gefaked is), stellen je hersenen zich meteen in op die persoon. De stemkloon hoeft dan niet eens meer heel goed te lijken omdat je hersenen eventuele onvolkomenheden automatisch afdoen als een slechte verbinding of omgevingsgeluid. Er is geen ruimte voor twijfel.
  Criminelen kunnen dit ook gebruiken om zich telefonisch bij een burger voor te doen als de vertrouwde (wijk)agent die vervolgens ‘collega’s’ langs stuurt, of als een bekende ‘collega’ die naar de politie belt om vertrouwelijke informatie te verkrijgen.

[36] https://www.anthropic.com/news/detecting-countering-misuse-aug-2025. Zie ook: https://www.trendmicro.com/en_us/research/25/i/evilai.html
[37] https://www.linkedin.com/pulse/unmasking-remote-fake-workers-ismael-alvarez-rkqxe geeft een goed inkijkje in hoe dit werkt en wat de motieven zijn. Bij deze vorm van criminaliteit kunnen ook geopolitieke motieven een rol spelen.
[38] https://www.trmlabs.com/resources/blog/ai-enabled-fraud-how-scammers-are-exploiting-generative-ai
[39] https://abcnews.go.com/GMA/News/dad-warns-ai-voice-scams-after-family-lost/story?id=99344931
[40] https://edition.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk
[41] https://edition.cnn.com/2023/04/29/us/ai-scam-calls-kidnapping-cec

- *Misleiding online biometrische authenticatiesystemen.*
  De kwaliteit van zowel gezichts- als stemklonen is inmiddels zo goed dat ze biometrische authenticatiesystemen voor de gek kunnen houden. Denk aan de Australische belastingdienst[42], waar je kon inloggen op basis van stemverificatie of online systemen voor het openen van een bank of cryptorekening[43] [44].

Alle informatie is gemanipuleerd

Er is nu veel aandacht voor malafide deepfakes en terecht want het wordt steeds makkelijker om bijvoorbeeld een gezicht of stem te klonen. Maar als je breder kijkt, zie je dat deepfakes onderdeel zijn van een grotere beweging waarbij alles synthetisch (nep) wordt[45]. Deze grotere beweging is veel belangrijker als het gaat om disruptie van processen, het tast fundamentele aannames aan.

Bij met AI gegenereerde of gemanipuleerde media gaat het meestal om toepassingen voor entertainment, dienstverlening of om onze kwaliteit van leven te verbeteren. Maar als politie gebruiken wij deze informatie bijna altijd in een andere context dan waar deze oorspronkelijk voor gegenereerd is. Aanvullingen in foto's door AI in gebruikte sociale media apps maken het moeilijk te achterhalen waar een foto genomen is. Dat AI in steeds meer systemen ingebakken zit en digitale content straks per definitie gemanipuleerd is maakt politiewerk veel complexer.

Een goed voorbeeld om dit te illustreren is videobellen, waarbij GenAI ook haar intrede heeft gedaan onder meer om bandbreedte te besparen[46]. Bijvoorbeeld door aan het begin van het gesprek een snapshot van je gezicht te maken, waarna tijdens het gesprek slechts enkele punten over de lijn gaan waarmee aan de ontvangende kant je gezicht door kunstmatige intelligentie opnieuw wordt gegenereerd.

Het is zelfs zo dat een aantal toepassingen er nu ook voor zorgt dat je elkaar op natuurlijke wijze in de ogen kijkt terwijl je misschien met hele andere dingen bezig bent[47] [48]. Allemaal nep dus, maar puur ter verbetering van de dienstverlening, want die gaat over het creëren van een vertrouwd dichtbij gevoel, alsof je echt fysiek tegenover elkaar zit en elkaar aankijkt in het gesprek.

---

[42]https://www.theguardian.com/technology/2023/mar/16/voice-system-used-to-verify-identity-by-centrelink-can-be-fooled-by-ai

[43] https://www.404media.co/inside-the-underground-site-where-ai-neural-networks-churns-out-fake-ids-onlyfake/

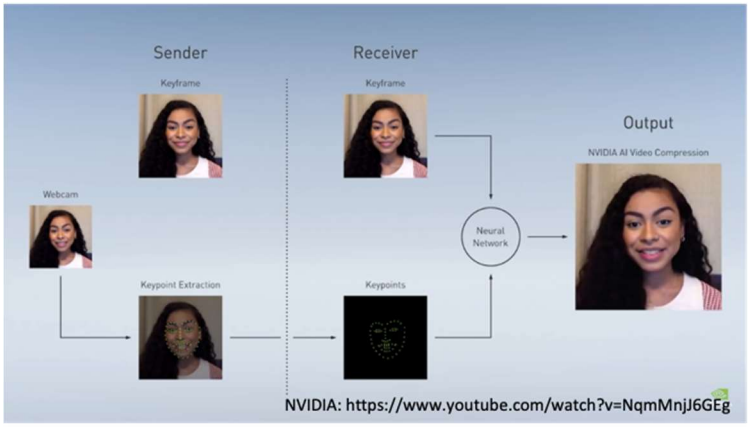
[44] Doordat in Nederland steeds meer banken bij het onboarden ook gebruik maken van de NFC chip in de identiteitsbewijzen, is een reguliere bankrekening openen niet of nauwelijks mogelijk. Wel kunnen online crypto en internationaal bankrekeningen worden geopend.

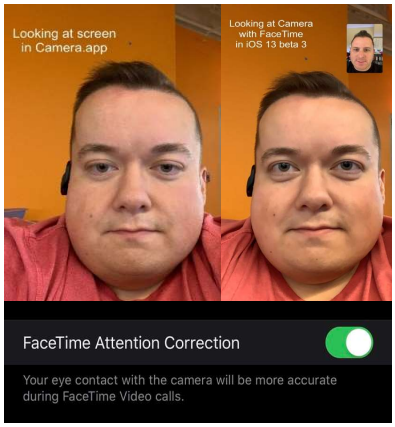
[45] https://www.bloomsbury.com/us/regulating-the-synthetic-society-9781509974948/, Zie ook Deepfakes: The Coming Infocalypse van Nina Schick (2020)

[46] https://www.youtube.com/watch?v=NqmMnjJ6GEg

[47]Facetime afbeelding komt van een inmiddels verwijderd account op X: https://twitter.com/WSig/status/1146146914985009154.

[48]https://appleinsider.com/articles/20/06/22/facetime-eye-contact-correction-feature-to-launch-with-ios-14 in een kleine steekproef bleek deze functie standaard aan te staan in de iphones.





Dit simpele voorbeeld laat goed zien hoe onze waarneming al wordt beïnvloed en het belang van context. De manipulatie in het videobellen is niet relevant in de context van een vergadering of Webinar. Maar we moeten het wel snappen als we met een slachtoffer of aangever in gesprek zijn. Of als dokter met een patiënt. Onbewust letten we dan immers ook op de non-verbale communicatie.

Maar wat is daarvan nog betrouwbaar als er maar een paar punten van het gezicht over de lijn gaan? Waarna op basis van de gemiddelden van een hele grote groep mensen, jouw gezicht in het beeld bij je gesprekspartner(s) gegenereerd wordt. Wat voor de massa geldt, hoeft niet op te gaan voor jou als individu.

Voordat we conclusies kunnen trekken naar aanleiding van digitale content of communicatie via de digitale snelweg is het van belang nader onderzoek te doen naar de betrouwbaarheid en bruikbaarheid in onze specifieke context.

Inzet capaciteit, wie heeft de regie

Naast dat we informatie gebruiken voor bewijs, is ook onze sturing erop gebaseerd. Informatie bepaalt hoe en waar wij onze capaciteit inzetten. Al eerder werd gerefereerd aan de inzet van capaciteit op het achterhalen van niet bestaande personen. Dit speelt ook in de context van kinderporno, waarbij de politie zo snel mogelijk de identiteit van het kind wil achterhalen om deze uit die situatie te halen. Door de toenemende kwaliteit van met GenAI gegenereerde beelden zal het steeds lastiger worden om vast te stellen waar het om een bestaand kind gaat.

GenAI kan worden ingezet om de openbare orde te verstoren, door bijvoorbeeld video's van nepaanslagen of door publieke figuren of betrouwbare bronnen dingen te laten zeggen en doen[49]. Ook kan het gebruikt worden om de chain-of-command te verstoren tijdens een crisis[50].

De afhandeling van klachten en verzoeken op basis van de Wet Open Overheid[51] zal ook meer capaciteit vragen. Mensen kunnen makkelijker brieven opstellen. Enerzijds zorgt dit dat meer mensen hun recht kunnen halen, maar diverse organisaties merken ook dat de verzoeken er veel

[49] In Baltimore werd een audio opname verspreid waarin een schoolhoofd zich racistisch uitliet. Deze deepfake leidde tot grote onrust. https://www.bbc.com/news/world-us-canada-68907895. In Nederland droegen, naast de filmpjes van echte mishandelingen, ook AI gegenereerde filmpjes van exploderende scholen bij aan onrust. https://nos.nl/artikel/2582100-middelbare-scholen-in-beverwijk-en-heemskerk-dicht-vanwege-jongerengeweld

[50] https://www.thejc.com/news/israel/israeli-secret-services-used-fake-phone-call-to-lure-irans-air-force-elite-to-their-deaths-ejqk7guy. Het is onduidelijk of hierbij ook een gekloonde stem is gebruikt, maar dit lijkt waarschijnlijk daar de orders meteen opgevolgd werden. Tijdens crises zullen mensen, vanwege de druk en het terugvallen op hiërarchie, hier sneller intrappen.

[51] Regelt in Nederland de openbaarheid van bestuur door openbaarmaking van informatie.

langer van worden, vol juridische, niet altijd kloppende, uitzettingen. En omdat de taalmodellen ook hierin hallucineren zijn organisaties veel meer tijd kwijt aan het analyseren van de brieven, ze zijn immers verplicht er adequaat op te reageren. Helaas biedt AI voor dit proces (nog) geen oplossing. Een andere zorg die in deze context naar boven komt is die van escalatie. Voorheen kon je bij klachten vaak nog met elkaar in gesprek om tot een oplossing te komen. De drempel om een advocaat te betrekken is immers hoog. Doordat het taalmodel een (schijnbare) juridische onderbouwing geeft voor de argumenten van de klager lijkt men minder snel bereid tot een verzoenend gesprek.

Een ander voorbeeld is het doen van valse meldingen, zoals zogenaamde noodsituaties, inclusief ondersteunende beelden. In Amerika hebben ze nu last van de Swattingservice[52] waarbij GenAI wordt ingezet om de SWAT unit (DSI) ergens op af te sturen. Deze service wordt bijvoorbeeld gebruikt door jongeren die bang zijn een toets niet te halen. Zij geven op deze website de naam van hun school door, betalen 75$ en de volgende ochtend wordt de lokale politie gebeld door bange (gegenereerde, synthetische) stemmen van niet bestaande kinderen die zich zogenaamd in de school verstopt hebben omdat er iemand met een wapen door de gang loopt.

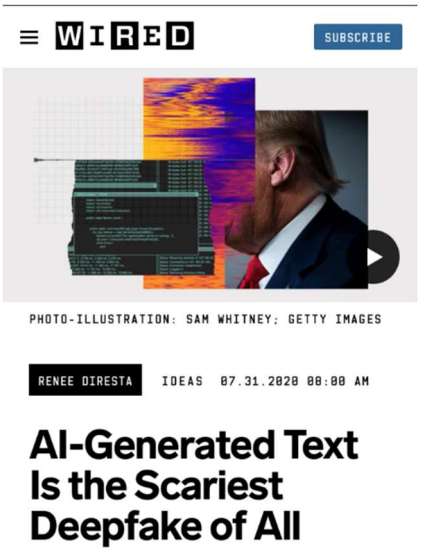
Ook op sociale media wordt steeds vaker AI ingezet om accounts of kanalen te vullen met steeds nieuw materiaal vanwege de advertentie-inkomsten. Een volledig door AI gegenereerd True Crime kanaal voedde complottheorieën over het falen van de mainstreammedia, omdat die geen aandacht gaven aan deze gewelddadige zaken. Ook de lokale politie werd erop aangesproken[53] waarbij het even duurde voordat deze begreep dat het om verzonnen casussen ging.

Het internet wordt overspoeld met AI gegenereerde content[54], niet alleen sociale media, ook verkoopsites als Amazon en Bol.com[55] lijden onder gegenereerde advertenties en gegenereerde boeken die zogenaamd door bekende schrijvers gepubliceerd zijn of over populaire onderwerpen gaan. Dat kan ook gevaarlijk zijn, zoals bij boeken over ADHD of boeken over het zelf paddenstoelen plukken waarin recepten met giftige paddenstoelen staan[56].

De sociale media platformen hebben belang bij deze content om mensen binnen hun omgeving te houden. Om voldoende advertentie-inkomsten te genereren is het noodzakelijk dat er steeds nieuwe content geplaatst wordt. Voor mensen is het lastig dagelijks nieuwe content te plaatsen, maar met generatieve AI kun je tientallen filmpjes per dag produceren en plaatsen. Bij vier van de top 10 meest populaire kanalen op Youtube is vastgesteld dat elke video GenAI materiaal bevat[57]. Ook de views en opmerkingen onder content zijn steeds vaker met AI gegenereerd. Dit maakt het voor de politie lastiger daar waar zij in haar werkprocessen gewend is op sociale media te vertrouwen voor bewijs- en sturingsinformatie.

[52] https://www.vice.com/en/article/torswats-computer-generated-ai-voice-swatting/
[53] https://www.404media.co/a-true-crime-documentary-series-has-millions-of-views-the-murders-are-all-ai-generated/
[54] https://www.404media.co/ai-slop-is-a-brute-force-attack-on-the-algorithms-that-control-reality/
[55] Het zoeken op “aangepaste titel” binnen Bol.com geeft op het moment van schrijven (juli 2025) veel relevante voorbeelden van met AI gegenereerde advertenties. Binnen de dropshipping community worden tutorials en tips hierover veel gedeeld.
[56] https://www.theguardian.com/technology/2025/may/04/dangerous-nonsense-ai-authored-books-about-adhd-for-sale-on-amazon https://www.404media.co/ai-generated-mushroom-foraging-books-amazon/
[57] https://sherwood.news/tech/ai-created-videos-are-quietly-taking-over-youtube/

De exponentiele groei van deze gegenereerde (onzin)content, ook wel AI Slop genoemd, leidt tot zichtbare disruptie van ons globale informatie ecosysteem[58]. Dat immers grotendeels afhankelijk is van het internet. Niet voor niks wees Wired in 2020[59] al op het gevaar van AI gegenereerde teksten. En toen was nog niet eens duidelijk dat binnen 5 jaar iedereen over elk willekeurig onderwerp, in elke taal, kwalitatief goede content zou kunnen genereren op maat gemaakt voor elke beoogde doelgroep. Waarbij kan niet alleen de productie van content, maar ook de distributie ervan volledig geautomatiseerd kan worden.



Als gevolg van AI Slop, mis- en desinformatie wordt het steeds lastiger om betrouwbare en relevante informatie te vinden, net zoals het steeds moeilijker wordt om jouw informatie bij anderen onder de aandacht te krijgen.

Desinformatie

Bij desinformatie gaat het om het opzettelijk verspreiden van misleidende informatie. De werking van ons brein in combinatie met de daarop afgestemde werking van veel bekende sociale media platformen en zoekmachines faciliteren zowel de snelle verspreiding als de impact van desinformatie.

Vaak wordt desinformatie zo verpakt dat het emoties oproept en de drang iets te willen doen, al was het maar een like, share of comment. Deze zogenaamde engagement is een belangrijke maatstaf voor het prioriteren van een bericht in nieuwsfeeds, tijdlijnen en zoekmachines.

Door niet alleen de productie en distributie uit te besteden aan AI, maar AI ook in te zetten om op jouw content te reageren verhoog je de engagement en komt jouw (des)informatie hoger in de tijdlijnen. Daarnaast wordt bij de verspreiding van desinformatie gebruik gemaakt van de volgende effecten:

- Majority Illusion
  De onderliggende structuur van sociale netwerken kan ervoor zorgen dat een gedraging of mening die wereldwijd zeldzaam is, systematisch oververtegenwoordigd kan zijn in een lokale omgeving, doordat bepaalde sleutelfiguren deze delen. Dit wordt “majority illusion” genoemd. Hierbij lijkt het alsof iedereen een bepaalde mening deelt, waardoor je geneigd bent erin mee te gaan. Hier wordt ook op ingezet met zogeheten ‘trollbots’ die in groten getale geautomatiseerd nepmeningen toevoegen aan discussies. Hierbij zie je dat steeds vaker deepfakes ingezet worden, zowel om de accounts authentieker te maken als de berichten die gedeeld worden;
- Illusory truth
  Dit begrip betreft onze neiging om valse informatie toch als echt te gaan zien als we er herhaaldelijk aan blootgesteld worden. Dit heeft te maken met onze hersenen die dingen die we zelf horen en zien vertrouwen. Zelfs al weten we dat iets niet klopt, gaan we er (onbewust) toch in geloven omdat we het zo vaak tegenkomen. Waardoor, zelfs als aangetoond is dat een beeld nep is, mensen toch komen met de reactie dat het beeld wellicht

[58] Zie ook het boek ‘Deepfakes: The Coming Infocalypse’ van Nina Schick (2020)
[59] https://www.wired.com/story/ai-generated-text-is-the-scariest-deepfake-of-all/

nep is, maar de boodschap erachter niet[60]. Dit maakt het ondermijnende effect van synthetische media extra groot. Wat we horen en zien is zo authentiek dat het moeilijk wordt om te blijven geloven dat het nep is.

Realistische audiovisuele media zijn steeds makkelijker met AI te genereren. Door de toenemende kwaliteit verlenen ze niet alleen extra authenticiteit, maar kunnen er ook meer emoties mee opgeroepen worden. Dit betekent dat nog meer mensen op een bericht zullen reageren of er aandacht voor hebben. GenAI is dus een sterk faciliterende factor, maar gelukkig komt desinformatie in Nederland nog weinig voor en is de impact nog laag[61].

4. Handelingsperspectief voor de politie

Voorgaande illustreert de uitdagingen waar we voor staan als politie, het wordt steeds lastiger om vast te stellen of informatie betrouwbaar is. Hieronder wordt ingegaan op een aantal oplossingsrichtingen voor met GenAI gegenereerde media.

Praktische handvatten

Eerder in dit artikel is onderscheid gemaakt tussen enerzijds specifieke GenAI, de traditionele deepfakes, en anderzijds generieke GenAI. De reden hiervoor is dat de huidige deepfake detectietools focussen op de traditionele deepfakes, ze zijn specifiek daarop getraind. Dit is waarom met andere tools gemanipuleerde afbeeldingen, ondanks dat ze voor ons duidelijk fake zijn, door de deepfake detectietools niet als zodanig worden herkend.

Vooral als het gaat om generieke GenAI zijn er nog praktische handvatten die iedereen kan gebruiken. De ontwikkelingen gaan snel[62], dus onderstaande tips zijn indicaties, geen bewijs! Daarnaast hebben ze betrekking op de pure resultaten van GenAI, waarbij geen nabewerking heeft plaatsgevonden!

GenAI algemeen:

- **1 = geen**. Generatieve AI maakt (genereert) beelden steeds vanaf nul, waardoor in verschillende afbeeldingen de manipulaties er verschillend uitzien. Vraag, bijvoorbeeld bij een aangifte of melding, om meer foto's of filmpjes, liefst genomen uit verschillende hoeken.
- Is het een buitenomgeving? Vergelijk via Google Streetview met de echte situatie.
- Kijk naar de eigenschappen (exif / metadata) van een bestand. Stel de beelden zelf veilig bij de bron, dan moeten de gegevens van de telefoon bij de eigenschappen staan. Na plaatsing op bijvoorbeeld sociale media of bij aanpassing door ChatGpt verdwijnen die.
- Kijk of je de foto op internet kunt terugvinden door, bijvoorbeeld via Google, te zoeken op afbeelding. Soms vind je dan dezelfde foto zonder manipulatie, ook kan de datum en context van die foto aanwijzingen geven.

[60] AI Slop: Last Week Tonight with John Oliver (HBO), https://www.youtube.com/watch?v=TWpg1RmzAbc
[61] https://decorrespondent.nl/15411/waarom-we-het-minder-over-desinformatie-moeten-hebben. Wat wel sterk groeit zijn AI gegenereerde filmpjes die bijvoorbeeld hitlerverheerlijking en geld
[62] Een actueel overzicht wordt bijgehouden op https://digitalks.eu/tips-voor-het-herkennen-van-genai

- AI is (voorlopig) slecht in natuurkundige wetten en daarom minder goed in kloppende schaduw of reflecties in ramen of plassen water, vaak zijn er in de achtergrond niet kloppende details te zien zoals lantaarnpalen die niet doorlopen, teksten die niet kloppen, of een perspectief dat niet juist is.
  De foto hiernaast laat dergelijke details duidelijk zien, zo is bijvoorbeeld de linkerhand van Trump te klein, loopt het mes onnatuurlijk onder het kopje door en klopt de vervorming van de lepel in het glas niet.
  Maar ook vanuit de context kan al meteen geconcludeerd worden dat de afbeelding nep moet zijn. Bij een koninklijk ontbijt zit je niet zo dicht op een gast, en een ei hoort niet thuis in het fruitschaaltje.

Deepfakes:

- Filmpjes of foto's waarin iemand iets doet of zegt dat diegene mogelijk niet gedaan heeft, worden meestal met traditionele deepfake tools gemaakt. De huidige kwaliteit is dermate goed dat je in de beelden zelf geen onvolkomenheden meer ziet. Technisch en contextgericht onderzoek zijn dan noodzakelijk, zoals zoeken of dezelfde beelden elders voorkomen, of er alternatieve informatie is (Osint), controleren of schaduwen & (weers)omstandigheden kloppen met geclaimde datum/tijdstip, kijken naar metadata en informeren in de omgeving of bij factcheckers.
- Bij live videoverbindingen gaat het tot nu toe altijd om de traditionele deepfakes. Vraag de persoon zijn camera (laptop/telefoon) in de hand te nemen en al bewegende in de ruimte rond te draaien. GenAI kan nog niet goed compenseren voor situaties waarin zowel de camera als de omgeving/persoon beweegt. Beter is natuurlijk dat je zelf het contact (opnieuw) initieert via het bij jou bekende telefoonnummer of email adres.

Deepfake detectietools, alleen voor specialisten als onderdeel van een bredere toolbox

Samen met de Universiteit van Amsterdam doet het Nederlands Forensisch Instituut (NFI) al jaren onderzoek naar deepfake detectietools[63]. Ook daaruit komt naar voren dat deze tools onbetrouwbaar zijn[64], en dat er grote behoeft is aan recente deepfakes die niet gebruikt zijn om

[63] https://www.forensischinstituut.nl/actueel/nieuws/2022/11/15/nieuwe-methodes-nfi-en-uva-voor-herkennen-deepfakes De besproken manuele methode gericht op visuele inspectie levert steeds minder op doordat de deepfaketools sterk verbeterd zijn. Inmiddels wordt per casus bekeken welke tools van meerwaarde kunnen zijn. In de praktijk blijft het een grote uitdaging, bijvoorbeeld als er alleen een screenshot van een afbeelding is, waardoor je ook geen metadata hebt.

[64] Nog steeds verschijnen er regelmatig onderzoeken waaruit zou moeten blijken dat een tool meer dan 90% van de deepfakes succesvol herkent. In praktijk blijken deze tools dan getest op verouderde datasets. Globaal kun je stellen dat deepfakes gemaakt met verouderde (vaak gratis) technologie inderdaad voor ruim 90% succesvol herkend kunnen worden. Bij testen op nieuwere deepfakes (max 2 tot 3 jaar oud) loopt dit percentage terug naar 50%. Dit betekent dat op 100 beelden er 50 ten onrechte als nep aangeduid kunnen worden. Dit maakt gebruik van deze tools door de politie onverantwoord, zij kan het zich niet permitteren om betrouwbare informatie opzij te schuiven als nep.

de tools te trainen, zodat het testen van tools onafhankelijk en op kwalitatief verantwoorde wijze gedaan kan worden.

In mei 2025 presenteerde het NFI hun onderzoek naar bloedstroomdetectie als mogelijke toekomstige aanvulling op de specialistische toolbox om deepfakes te detecteren. Deze methode moet nog wel worden gevalideerd[65]. Hopelijk komt dit op tijd, nu onderzoek heeft aangetoond dat recente deepfakes het hartslagpatroon uit het oorspronkelijke materiaal overnemen. Daarmee vervalt de methode die uitging van de subtiele volumeveranderingen van een bloedvat die ontstaan door de hartslag[66].

Zelfs als ze goed zouden zijn, hebben de detectietools maar beperkt meerwaarde. Ze zijn getraind op specifieke GenAI terwijl de toename in synthetische media vooral komt van generieke GenAI. En als ze daarvoor zouden werken, dan zouden ze nagenoeg altijd moeten aanslaan nu overal GenAI wordt toegepast. Het onderscheid tussen nep en echt wordt daarmee irrelevant. De term "nep" heeft eigenlijk geen waarde meer, waar het om draait is de betrouwbaarheid en toepasbaarheid van de betreffende informatie in een specifieke context. Denk aan de AI-filters op sociale media als snapchat, in de meeste gevallen zijn deze niet relevant. Maar wellicht dat bepaalde filters er wel toe kunnen leiden dat een persoon niet herkend wordt wanneer je de afbeelding gebruikt voor gelaatsvergelijking.

Ook met AI gemanipuleerde informatie kan betrouwbaar en relevant zijn. Om dit vast te stellen is het belangrijk te begrijpen welke manipulaties er hebben plaatsgevonden. Dat is waar een combinatie van technische tools wel degelijk kan bijdragen. Het NFI werkt volgens deze methodiek, zij gebruiken een diverse combinatie van, soms zeer specialistische, technische tools met elk hun eigen onzekerheden, om te komen tot een uitspraak over de betrouwbaarheid. Het vergt nog een flinke investering in kennis en kunde[67] voor de politie deze aanpak intern kan inbedden. En uiteraard geldt dit net zo goed voor het Openbaar Ministerie en de Rechtspraak.

Regulering

Volgens de Europese AI Act[68] moet AI-gegenereerde content als zodanig worden gelabeld door de platformen. Daarnaast moeten ontwikkelaars van (tools voor) synthetische media zorgen dat deze gedetecteerd kunnen worden, bijvoorbeeld door watermerken.

Het is al duidelijk dat de platformen niet voldoen aan de labelings-eis. Bijvoorbeeld bij de eerder genoemde populaire GenAI-kanalen op Youtube wordt niet altijd vermeld dat AI wordt gebruikt. Ook staat het internet vol met tutorials over hoe je watermerken kunt verwijderen uit GenAI content.

De vraag is niet of de EU voldoende streng gaat handhaven om te zorgen dat er wel aan de regulering voldaan wordt. De echte vraag is wat de meerwaarde van dergelijke regulering nog is als nagenoeg alle content straks op enigerlei wijze door AI gegenereerd of gemanipuleerd is.

Ook hier speelt dat het onderscheid tussen wel of niet met AI gemanipuleerd irrelevant wordt als het gebruik zo wijdverbreid is en GenAI standaard in allerlei toepassingen, zoals onze

[65] https://www.forensischinstituut.nl/actueel/achtergrondverhalen-nfi/hoe-onze-hartslag-kan-verraden-dat-een-video-deepfake-is---eafs-2025-deel-2

[66] https://www.frontiersin.org/journals/imaging/articles/10.3389/fimag.2025.1504551/full

[67] Dit gaat bijvoorbeeld om het kunnen omgaan met likelyhood ratio's, wat betekent het als iets voor 60% waarschijnlijk is. En hoe combineer je dat met een waarschijnlijkheidspercentage van 40% uit een andere tool. Wat betekent de stapeling van aannames etc.

[68] https://www.technologyreview.com/2024/03/19/1089919/the-ai-act-is-done-heres-what-will-and-wont-change/

telefooncamera's, is geïntegreerd. De echte behoefte zit in het laagdrempelig kunnen duiden van de betrouwbaarheid en het tegengaan van malafide toepassingen.

Provenance: grip op betrouwbaarheid voor het grote publiek

Niet alleen binnen waarheidsvinding is het belangrijk om de betrouwbaarheid van informatie te duiden, ook bij bijvoorbeeld desinformatie speelt dit een grote rol. Provenance gaat verder dan simpel labelen of watermerken. Het Engelse begrip Provenance[69] laat zich vertalen als herkomst, maar de betekenis in deze context is breder. Het gaat om het met behulp van cryptografisch versleuteling digitaal onherroepelijk en onweerlegbaar vastleggen van bewijzen van de herkomst, toegebrachte wijzigingen of juist de onveranderlijkheid van afbeeldingen en bijbehorende metadata. Het betreft dus niet alleen het op het moment van creatie al vastleggen en meegeven van de herkomst van informatie en de processen en methodologie waarmee het is geproduceerd. Het gaat ook om het inzichtelijk maken en kunnen controleren van alle tussenstappen tot de digitale content uiteindelijk bij de consument (de uiteindelijke afnemer van de content) terecht komt.

Het mooie van provenance is enerzijds dat je als politie of andere organisatie het mogelijk maakt voor mensen om laagdrempelig te checken of de informatie echt bij jou vandaan komt en tussentijds niet gemanipuleerd is. Dit draagt bij aan vertrouwen in de politie. Anderzijds voorkom je discussies over wat wel of niet desinformatie is. Mensen kunnen altijd toetsen of de informatie echt van de betreffend de bron komt, zoals de NOS, maar hebben vervolgens de vrijheid om zelf te bepalen of ze deze bron vertrouwen.

Het internationale Content Authenticity Initiative (CAI)[70], waarbij partijen als Microsoft, Adobe, de BBC, Samsung, Canon en Reuters zijn aangesloten, heeft al diverse tools en standaarden hiervoor opgeleverd. In Nederland is adoptie helaas nog niet van de grond gekomen.

Tegengaan malafide gebruik

Handhaving en regulering zijn vooral belangrijk als het gaat om applicaties die het gebruik van GenAI in malafide zin niet tegengaan of zelfs aanmoedigen. Denk aan Grok, het door Elon Musk op de markt gebrachte alternatief voor ChatGPT dat geen begrenzingen kent, waardoor er bijvoorbeeld gemakkelijk naaktbeelden mee gemaakt kunnen worden[71]. Maar ook de diverse deepporn en uitkleed(nudify)apps.

Met de huidige wetgeving kan hier nu al tegen opgetreden worden. Maar dergelijke apps staan nog gewoon in de Appstore van onder meer Apple en worden ze pas verwijderd als journalisten er aandacht aan besteden[72]. Actieve en dwingende handhaving hebben echt effect zoals Apple[73]

[69] Op 26 januari 2023 vond de eerste rijksbrede werkconferentie plaats over de impact van synthetische media & deepfakes. Het verslag en de greenpaper over provenance zijn hier gepubliceerd: https://digitalks.eu/VerslagWerkconferentieImpactSynthetischeMedia

[70] https://contentauthenticity.org/

[71] https://www.theverge.com/report/718975/xai-grok-imagine-taylor-swifty-deepfake-nudes

[72] https://www.404media.co/congress-pushes-apple-to-remove-deepfake-apps-after-404-media-investigation/

[73] https://www.theverge.com/2018/2/5/16974710/apple-telegram-ios-app-store-removal-explanation-child-pornography-distribution

en Frankrijk[74] eerder lieten zien in hun aanpak van Telegram. De Nederlandse en Belgische overheid zou hierin meer lef en initiatief kunnen tonen.

In 2021 heeft het WODC[75] onderzocht of de bestaande wetgeving voldeed in het licht van "*huidige en toekomstige onrechtmatige of strafwaardige uitingsvormen van deepfaketechnologie".* Daaruit bleek dat vrijwel alle bepalingen in het wetboek van strafrecht voldoende technologie-neutraal zijn geformuleerd. Op twee punten werden aanpassingen in overweging gegeven. Met de introductie van Art.254baSr is inmiddels ook het vervaardigen en bezit van pornografische deepfakes strafbaar (als daar geen toestemming voor is gegeven).

De andere suggestie betrof een lacune in artikel 231a en bSr. Omdat hierin alleen het misbruik van biometrische gegevens in de context van identificatie strafbaar is gesteld. Dat maakt dat je als burger afhankelijk bent van de interpretatie van de rechter van 231bSr wanneer je op een andere manier benadeeld bent[76]. De suggestie om het artikel breder te trekken is helaas niet overgenomen.

In deze context heeft Denemarken ervoor gekozen om mensen copyright te geven op hun eigen lichaam, gezichtskenmerken en stem. Hiermee krijgen Denen het recht om van platformen te eisen dat ze materiaal verwijderen dat zonder toestemming is geplaatst[77].

Maar, zoals betoogd door Etienne Valk van het Instituut voor Informatierecht van de Universiteit Utrecht[78], het probleem van auteursrecht is dat dit overdraagbaar is. Met een aantrekkelijk aanbod van Big Tech of anderen kunnen mensen verleid worden tot het overdragen van de rechten waardoor ze alle grip kwijt zijn[79]. Uitbreiding van het portretrecht is een stuk logischer omdat het vergaand overdragen van zeggenschap bij het portretrecht niet mogelijk is. Daarnaast, zoals Valk aangeeft: *"Het is simpelweg nodig om niet, zoals nu, alleen het gezicht van een persoon eronder te laten vallen, maar ook diens stem en lichaam. Doop het om tot „persoonskenmerkenrecht", bijvoorbeeld."* Deze verbreding is echt noodzakelijk omdat

---

[74] https://apnews.com/article/france-telegram-pavel-durov-arrest-6e213d227458f330ed16e7fe221a696c
https://www.wired.com/story/pavel-durov-arrest-telegram-content-moderation/

[75] https://repository.wodc.nl/handle/20.500.12832/3134

[76] art. 231a lid 1 Sr: 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.
art. 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.
Deepfakes betreffen het vervalsen van beeld (gezicht) en audio (stem) – dat zijn bij uitstek biometrische persoonsgegevens. Maar inmiddels zijn er een aantal strafzaken waarin toch succesvol een beroep gedaan is op 231b, onder meer door de rechtbank den Haag die concludeerde dat: "de foto's waarop het hoofd en een deel van het bovenlichaam van het slachtoffer was te zien, wél kunnen worden aangemerkt als "identificerende persoonsgegevens, niet zijnde biometrische gegevens" en dat daarmee voldaan is aan de kwalificatie van art. 231b Sr. ECLI:NL:RBDHA:2018:15097(Rechtbank den Haag) en 23-000843-19 (Hoge Raad)

[77] https://nos.nl/nieuwsuur/artikel/2574932-in-de-strijd-tegen-deepfakes-krijgen-denen-copyright-op-eigen-gezicht-en-stem
https://www.theguardian.com/technology/2025/jun/27/deepfakes-denmark-copyright-law-artificial-intelligence

[78] https://www.nrc.nl/nieuws/2025/07/29/optreden-tegen-deepfakes-is-een-goed-idee-maar-doe-dat-dan-wel-via-het-portretrecht-a4901612

[79] Denk aan acteurs die zich gedwongen voelen om zich te laten scannen;
https://www.businessinsider.com/background-actor-extra-body-scans-hollywood-ai-fears-report-2023-8
https://www.npr.org/2023/08/02/1190605685/movie-extras-worry-theyll-be-replaced-by-ai-hollywood-is-already-doing-body-scan Mensen kunnen de gevolgen niet overzien bij verkoop van hun digital twin voor reclame doeleinden:
https://www.nytimes.com/2025/08/17/business/tiktok-ai-avatars.html?unlocked_article_code=1.fE8.hv1i.Tyof97BfHZ7X&smid=url-share

.

steeds meer (gedragskenmerken) gemeten, geanalyseerd en gebruikt worden, ook in de context van identificatie.

In alle gevallen blijft het wel van belang dat de overheid dan wel handhavend optreedt.

Als het gaat om het voldoen van de wetgeving zijn er sinds het rapport van WODC echter ook nieuwe vraagstukken ontstaan. Denk aan de Snapchat-casus over grooming, de eerder genoemde AI-chatbot van Meta die proactief sexueel getinte gesprekken mag voeren met kinderen en (therapie)bots die volwassenen aanmoedigen in ongewenst en soms strafbaar gedrag. In hoeverre kan hier vanuit wet- en regelgeving tegen opgetreden worden?

Het verkrijgen van meer zekerheid als professionele standaard

Voor de politie is de grootste winst te behalen in de werkprocessen en in de kennis en vaardigheden van medewerkers en leidinggevenden. Dit gaat over bewustwording, en het vermogen tot kritisch reflecteren en regelmatige ruggespraak met collega's en experts binnen en buiten de politie.

Allereerst moeten we ons ervan bewust worden dat veel van onze aannames, gebaseerd op jarenlange ervaring, niet langer kloppen. Zoals de zelfmoordcasus, waarbij nog aangenomen werd dat een stem alleen door de persoon zelf gebruikt kan worden. Verder is het ook belangrijk te snappen dat we zelf ook beïnvloed kunnen worden door desinformatie, ook onze hersenen zijn geneigd dat wat we zelf horen en zien automatisch te vertrouwen. Ook wij kunnen gebeld worden door iemand die de stem van een vertrouwde collega misbruikt, of te snel conclusies trekken als we een stem denken te herkennen in een onderzoek. En als laatste het bewustzijn dat ook als er geen malafide intenties spelen, we toch moeten kijken naar manipulaties, omdat deze relevant kunnen zijn in de context van ons onderzoek.

Regelmatig zal dit aanleiding geven tot extra onderzoek, om zo meer zekerheid te krijgen over de betrouwbaarheid. De eerder genoemde handvatten zullen hierbij helpen, net als het begin 2025 opgeleverde handelingskader over hoe om te gaan met mogelijk AI-gegenereerde of gemanipuleerde stemmen.

In dat handelingskader is ook meegenomen dat het in veruit de meeste zaken voor de strafbaarstelling niet uitmaakt of het om een deepfake gaat, zoals bij oplichting en afpersing. Het is wel relevant als het gaat om het vaststellen van de identiteit; heeft deze persoon inderdaad ingelogd bij de belastingdienst, was die verdachte inderdaad betrokken in het afgeluisterde gesprek, of is dit inderdaad het ontvoerde slachtoffer.

Het handelingskader laat zien dat al snel een specialist van Digitale Opsporing betrokken moet worden, en wanneer deze specialisten moeten schakelen met het NFI. Als het gaat om het duiden van de betrouwbaarheid van beelden zal ook geschakeld moeten worden naar digitaal-forensisch specialisten voor onderzoek naar metadata en apparatuur, of Osint specialisten voor verdieping op het internet.

Het is duidelijk dat de vraag naar expertise in de 2^{e} en 3^{e} lijn (NFI) hierdoor zal groeien.

Transparantie over gemaakte afwegingen

Om kwalitatief goede beslissingen te nemen is het van belang om voldoende zekerheid te hebben over de betrouwbaarheid van de informatie waarmee wordt gewerkt. Hoe groter de impact van het te nemen besluit, hoe meer zekerheid nodig is. Als er nu beelden aangeleverd

worden met daarin iemand die een strafbaar feit pleegt, dan kun je je afvragen of de beelden gemanipuleerd zijn. Gaat het om beelden die buren van elkaar gemaakt hebben, dan kun je bij de wijkagent informeren of er iets speelt tussen deze partijen. Krijg je geen extra zekerheid over de (on)betrouwbaarheid, dan is het wellicht beter de betreffende persoon aan te lopen of uit te nodigen op het bureau in plaats van om vijf uur 's ochtend de deur in te trappen.

We werken continu met onzekerheid, het is vaak lastig vast te stellen wat nu echt waar is, maar je kunt dus wel een bepaald scenario plausibeler achten dan andere scenario's. Gaat het om camerabeelden van een winkel dan kun je je ook afvragen of iemand de gelegenheid had om de beelden te manipuleren en of het technisch aannemelijk is dat de beelden gemanipuleerd zijn. Zijn er meerdere camera's die beelden van dezelfde situatie uit verschillende hoeken hebben genomen? Hoe plausibel is dat meerdere camerabeelden tegelijkertijd zijn gemanipuleerd[80]?

Deze reflectie (afwegingen) en de eraan gerelateerde onderzoekstappen moet transparant plaatsvinden, enerzijds om als instituut toetsbaar en betrouwbaar te blijven, anderzijds om hele praktische redenen. In de rechtszaal kan de verdediging immers claimen dat bewijs niet authentiek is. Dit deepfake verweer komen we al tegen in de rechtszaal[81]. Door in het onderzoek al te kijken naar mogelijke manipulaties en in het proces-verbaal toe te lichten dat dit gedaan is en hoe dit van invloed is op de bewijswaarde, wordt de kwaliteit verhoogd. Daarnaast bespaart het capaciteit omdat de politiecollega's zich niet opnieuw hoeven in te lezen in de zaak om achteraf, weken of maanden na indiening, nog vragen te beantwoorden van de rechter.

Naast deze deepfake verdediging is er ook het risico van introductie van informatie als bewijs, waarbij het om onbetrouwbaar materiaal gaat. Dit gaat niet alleen om informatie van het internet. Het kan ook gaan om informatie van bijvoorbeeld (keten)partners of door onszelf gegenereerde informatie. In Nederland is het bijvoorbeeld nog steeds toegestaan om zelf een pasfoto aan te leveren voor je paspoort of rijbewijs. Met de wetenschap dat het bedrijf McAfee al in 2020 een pasfoto wist te genereren die voor de mens op persoon A en voor een gezichtsherkenningssysteem op persoon B lijkt[82], rijst de vraag in hoeverre we als politie kunnen vertrouwen op deze bij gemeenten op te vragen pasfoto's.

5. Conclusie: investeren in kritisch vermogen, reflectie en expertise

We bevinden ons op een kantelpunt. GenAI heeft, haast ongemerkt, zijn intrede gedaan in ons dagelijks leven via onze telefoons, online platformen, apps, camera- en communicatiesystemen, slimme apparaten thuis en voertuigen.

Hoewel GenAI criminaliteit en desinformatie efficiënter en effectiever kan maken, is de grootste impact voor de politie, als informatie-intensieve organisatie, de verstoring van het informatie-ecosysteem.

Hoewel zelden malafide, beïnvloeden met AI gegenereerde manipulaties namelijk wel onze perceptie. Ze worden gecreëerd binnen specifieke contexten, zoals dienstverlening en entertainment, die verschillen van de context waarin wij, als politie, de betreffende informatie willen gebruiken. Dit kan de waarde van bewijsmateriaal en besluitvorming aantasten.

[80] Gelukkig nog niet!

[81] Bijvoorbeeld zaken waarin verdachten claimen dat ze niet zelf een bankrekening geopend hebben, hun verweer is dat dit een deepfake moet zijn geweest. Het advies is nu om altijd de oorspronkelijke beelden bij de bank op te vragen en al in het Proces Verbaal te vermelden dat dit onderzoek is gedaan en wat daaruit kwam over de betrouwbaarheid.

[82] https://www.technologyreview.com/2020/08/05/1006008/ai-face-recognition-hack-misidentifies-person

Daarom is het essentieel om zowel meer zekerheid over de betrouwbaarheid van informatie te verkrijgen als om met de toenemende onzekerheid om te gaan. Alleen dan kan de politie de kwaliteit van haar besluitvorming waarborgen in de context van waarheidsvinding en een efficiënte en effectieve inzet van mensen en middelen.

Het advies aan de politie is om:

1. De overheid te stimuleren tot het tegengaan en voorkomen van malafide gebruik van A.I.;
    a. Een veel strengere handhaving van wet & regelgeving;
    b. Uitbreiden portretrecht met alle biometrische (gedrags)kenmerken
    c. Onderzoeken hoe platformen en aanbieders van GenAI aangepakt kunnen worden, zodat deze voldoen aan regelgeving voor ze in de markt gezet worden en niet zoals nu pas op basis van incidenten in de praktijk (soms en maar deels) worden aangepast.
    d. Waar nodig wet- en regelgeving hierop aanpassen.
2. Burgers handvatten te geven om de betrouwbaarheid van (politie)informatie te controleren door provenance te integreren in alle door de overheid en politie gegenereerde, geproduceerde en gedeelde informatie.
3. Binnen de politieorganisatie en haar ketenpartners te investeren in:
    a. Brede bewustwording en kennis van handelingsperspectief;
    b. Leiderschap dat kritische reflectie en het ter discussie stellen van aannames stimuleert door hier zelf een voorbeeldrol in te pakken en collega's hier zichtbaar op te bevragen;
    c. Ondersteunende expertises in de tweede en derde lijn;
    d. Het bijhouden van ontwikkelingen;
4. Wetenschappelijke instituties te stimuleren om nader onderzoek te doen naar zowel de sociaal-maatschappelijke effecten van gebruik en misbruik van A.I., als ook technologieën die bijdragen aan het waarborgen van authenticiteit en betrouwbaarheid van informatie.
5. Samen met NFI en Europese politiekorpsen bouwen aan een testset met recente deepfakes ten behoeve van het onafhankelijk kunnen testen van tools.