EFCSN

Hoe je AI-gegenereerde of digitaal bewerkte content herkent

Toegankelijke AI-modellen kunnen mensen helpen bij het leren over en maken van content. Maar ze kunnen ook **de risico's van mis- en desinformatie op onze open samenleving en het democratisch debat vergroten**. Het is belangrijk om te voorkomen dat onze gedeelde informatieruimtes vol komen te staan met door AI-gegenereerde en digitaal bewerkte mis- en desinformatie.

Deel van de oplossing zijn nieuwe technologieën, die de herkomst van content kunnen bepalen of detectiesoftware. **Deze technische oplossingen zijn alleen niet perfect. We hebben onafhankelijke factcheckers nodig om de samenleving te voorzien van een gemeenschappelijke basis in geverifieerde feiten.**

Hier volgt een overzicht hoe professionele en onafhankelijke factcheckers AI gegenereerde mis- en desinformatie herkennen en ontkrachten - en wat jij van ze kunt leren.

AI-gegenereerde content neemt toe

Momenteel vormt AI-gegenereerde desinformatie maar een klein deel van alle claims die door professionele en onafhankelijke factcheckers worden onderzocht. Digitaal bewerkte inhoud komt een factchecker vaker tegen.

Maar uit een intern onderzoek onder EFCSN-leden blijkt dat **de meeste factcheckers het erover eens zijn dat AI-gegenereerde en digitaal bewerkte content steeds belangrijker wordt**. Recente voorbeelden met betrekking tot de Europese verkiezingen bevestigen dit.

KORT UITGELEGD: *Digitaal bewerkt* is inhoud die gewijzigd is om de oorspronkelijke boodschap te veranderen. AI-tools kunnen hiervoor ingezet worden. Bewerkingen die iets verduidelijken of de kwaliteit verbeteren vallen hier niet onder.

AI-gegenereerd verwijst naar elke vorm van content die is gemaakt door artificiële intelligentie.

Deze informatie wordt aangeboden door **AI@EUElections,** gefinancierd door Meta & beheerd door het **European Fact-Checking Standards Network** in samenwerking met haar leden en externe experts.

Technologie ontwikkelt zich snel, maar alleen daarop kunnen we niet vertrouwen.

AI-experten en professionele factcheckers zijn het eens: **AI-detectietools alleen zijn niet genoeg om AI-gegenereerde of digitaal bewerkte content te ontkrachten.**

In plaats van een AI-detector te gebruiken, raden experts aan allereerst te begrijpen hoe AI-modellen werken, en waar ze hun inhoud vandaan halen. Samen met wat hulp van statistieken, kunnen factcheckers zo de sterke en zwakke kanten van AI-tools herkennen - en diens kans op succes. **Maar toch, ook bepaalde tools kunnen een goed uitgangspunt zijn.**

Initiatieven die **de herkomst van content kunnen herkennen**, zoals C2PA, helpen de bron en bewerkingsgeschiedenis te achterhalen. Zulke watermerken en controles zijn alleen niet onfeilbaar.

HOE AI-DESINFORMATIE MENSEN BEÏNVLOEDT

"Elke keer dat je vanuit je onderbuikgevoel reageert, ga je voorbij aan het denkproces."
– Christine Dugoin*

PSYCHOLOGIE:
Beïnvloedingsoperaties zijn vaak ontworpen om onze psychologische vooroordelen uit te buiten.

Inzicht in je eigen vooroordelen, en die van je publiek, kan helpen om desinformatie tegen te gaan.

DOELEN: Waarom zou iemand met kwade bedoelingen AI gebruiken om desinformatie te creëren en verspreiden? Welk doel heeft dat in de echte wereld?

- ➔ Hun bereik uitbreiden naar een ander land of andere gemeenschap?
- ➔ Onopgemerkt blijven of juist factcheckers willen overspoelen met variaties op dezelfde bewering?
- ➔ Meningen beïnvloeden, door een netwerk van neppe accounts dat geloofwaardig lijkt?

** Christine Dugoin is onderzoeker naar de invloed van informatie aan La Sorbonne.*

Ontkrachten vraagt om een veelzijdige aanpak en genuanceerd begrip.

Als detectietools niet werken, wat dan wel? Het is net zo belangrijk om de context van een bewering te begrijpen als de inhoud. Professionele factcheckers zijn experts in de nodige onderzoeks-vaardigheden. Hier een paar tips.

> *"Detectietools zullen nooit voor 100% werken - ik verwacht niet dat ze dat ooit zullen doen"*
> - Henk van Ess**

GA DE BRON NA: Kun je de herkomst vaststellen? Waar gaat de bron over, en wat wordt er gedeeld? Wie reageert op de content? Welk effect zou deze content kunnen hebben op lezers?

BETROUWBAARHEID VASTSTELLEN: Verifieer de informatie, onafhankelijk en bij geloofwaardige bronnen - zoals bij experts in het specifieke onderwerp. Gebruik je eigen kennis: is dat wat je ziet wel logisch?

Gebruik **FORENSISCHE MEDIA** technieken als aanvulling op traditionele onderzoeksvaardigheden. Denk aan: data-scraping, geolocatie, biometrische gezichtsherkenning, patroon analyse, etc.

LEER & ONTWIKKEL: Makers van AI-gegenereerde desinformatie passen zich voortdurend aan. Beweeg in je werkwijze dus altijd mee met het veranderende landschap.

DEEL JE WERK

Experts raden aan om, naast de ontkrachte bewering, ook een heldere analyse en links naar de gebruikte bronnen te geven. Dat helpt lezers een onderzoek te volgen en een genuanceerd verhaal te begrijpen. In sommige gevallen is het onderzoek belangrijker dan het feit dat de content door AI gemaakt is.

*** Henk van Ess is een expert in OSINT en factchecking-technieken*

EFCSN

Beknopte handleiding: *Dos, Don'ts* en hints

Hieronder staan hints waarmee je kunt herkennen of content door AI-gegenereerd of digitaal bewerkt is. Samen met andere tips in deze gids (context, onderzoekstechnieken en detectietools) kunnen ze je helpen om de waarheid te doorgronden achter dat wat je ziet.

Tekst

- Vaak (maar niet altijd) **betere spelling** dan een mens.
- Gebruik van **te formele taal,** vooral voor de context op sociale media.
- **Te veel bijwoorden en bijvoeglijke naamwoorden.**
- Gebrek aan menselijke emotie, humor, sarcasme en idiomatische uitdrukkingen.
- Mogelijk **gebrek aan specifieke details** (namen, data, locaties) of originele ideeën.
- En het belangrijkst: kloppen de feiten in de tekst?

Video

- Gebruik geen videostill in een AI-gegenereerde beelddetector.
- Kijk naar **gezichtsuitdrukkingen en bewegingen**, zoals knipperen of het bewegen van de mond bij de audio.
- Bevatten soms **harde overgangen of cuts**.

Audio

- **Vergelijk verdachte audio met een originele sample** met behulp van tools die verschillen kunnen opsporen in spraak- en adempatronen, intonatie...
- Vermijd geluidsfragmenten van lage kwaliteit als je een geluidsdetector gebruikt.
- Kan gekenmerkt worden door een **onnatuurlijke manier van spreken**, gebrek aan pauzes of natuurlijke ademhaling.

Afbeeldingen

- Kijk naar details die onnatuurlijk zijn: perfecte huid, wazige achtergrond, onnatuurlijke schoonheid of licht en eigenaardigheden zoals extra vingers.
- **Zoek naar een watermerk** van gewone afbeelding-generators.
- Let op de details: zijn die logisch? Is het gepast?
- Als je een detector inzet, gebruik dan **een afbeelding met hoge resolutie, of een van de eerste geüploade versies**. En dus niet de veel gedeelde afbeelding.