EFCSN

Cum să identifici și să înțelegi conținutul generat de AI sau alterat digital

Modelele de inteligență artificială accesibile și ușor de utilizat pot ajuta oamenii să învețe și să creeze conținut, dar pot, de asemenea, **amplifica riscurile pe care le reprezintă dezinformarea** pentru societățile deschise și discursul democratic. Este important să prevenim ca spațiile noastre comune de informare să nu fie aglomerate cu dezinformări generate de AI și modificate digital.

O parte a soluției este reprezentată de noile tehnologii, cum ar fi inițiativele de detectare a proveniență conținutului și software-ul de detectare. **Dar soluțiile tehnologice sunt departe de a fi perfecte și avem nevoie de activitatea unor fact-checkeri independenți pentru a oferi societății un set comun de fapte verificate.**

Iată o scurtă trecere în revistă a ceea ce fac fact-checkerii profesioniști și independenți pentru a identifica și demitiza dezinformările generate de AI și ce puteți învăța de la ei.

Conținutul generat de AI este în creștere

În prezent, dezinformarea generată de AI reprezintă un mic procent din toate afirmațiile investigate de fact-checkeri profesioniști și independenți. Conținutul alterat digital este mai frecvent în activitatea fact-checkerilor.

Dar: într-un sondaj intern realizat în rândul membrilor EFCSN, **majoritatea fact-checkerilor au fost de acord că, în viitor, conținutul generat de AI și modificat digital va deveni din ce în ce mai relevant**. Iar exemplele recente din contextul alegerilor europene susțin acest pronostic.

ÎNȚELEGEȚI: *Modificat Digital* se referă la orice formă de conținut care a fost modificat semnificativ pentru a manipula sau schimba mesajul transmis inițial, inclusiv modificările efectuate cu instrumentele de AI. Asta nu include editarea pentru claritate sau calitate.

Generat de AI se referă la orice formă de conținut care a fost creat de un sistem de inteligență artificială.

Informații oferite de **AI@EUElections**, finanțate de Meta și gestionate de **European Fact-Checking Standards Network** în colaborare cu membrii săi și experți externi.

Tehnologia evoluează rapid, dar nu ne putem baza doar pe ea.

Experții în AI și fact-checkerii profesioniști sunt de acord: **Instrumentele de detectare a AI nu sunt suficiente pentru a detecta sau demitiza conținutul generat de AI sau modificat digital.**

Înainte de a utiliza un detector, experții recomandă să vă familiarizați cu generatoarele și detectoarele de conținut AI. Înțelegând modul în care sunt antrenate modelele și cunoscând un pic de statistică, fact-checkerii pot începe să recunoască punctele forte și punctele slabe ale unui instrument, precum și probabilitatea de succes a acestuia. **Chiar și așa, instrumentele pot fi un punct de plecare util.**

Inițiativele de **proveniență a conținutului**, cum ar fi specificațiile C2PA, pot contribui la certificarea sursei și a istoricului conținutului media, dar filigranarea și verificarea nu sunt ireproșabile.

Cum afectează dezinformarea AI oamenii

„De fiecare dată când reacționezi, dacă pot spune așa, cu curajul tău, acesta îți ocolește reflecția."
– Christine Dugoin*

PSIHOLOGIE: Operațiunile de influență sunt adesea concepute pentru a profita de prejudecățile psihologice.

Înțelegerea propriilor prejudecăți și a celor ale publicului dumneavoastră poate contribui la combaterea dezinformării.

OBIECTIVE: De ce s-ar putea ca un actor rău să se bazeze pe AI pentru a crea sau răspândi dezinformare? Care este impactul pe care intenționează să îl aibă în lumea reală?

- ➔ Să-și extindă raza de acțiune într-o altă țară sau comunitate?
- ➔ Să evite detectarea sau să copleșească fact-checkerii prin generarea mai multor variante de afirmații similare?
- ➔ Să influențeze gândurile sau convingerile prin stabilirea credibilității prin intermediul unor rețele de conturi neautentice?

** Christine Dugoin este cercetător în domeniul influenței informaționale la Sorbona.*

Demontarea necesită o abordare cu multiple fațete și o înțelegere nuanțată

Deci, dacă instrumentele de detectare nu funcționează, atunci ce funcționează? Este important să înțelegem contextul unei afirmații la fel de mult ca și conținutul acesteia. Fact-checkerii profesioniști sunt experți în competențele de investigare necesare. Iată câteva sfaturi.

> *"Instrumentele de detectare nu vor funcționa niciodată 100% - nici nu mă aștept să o facă vreodată."*
> - Henk van Ess**

LUAȚI ÎN CONSIDERARE SURSA: Puteți confirma identitatea lor? Despre ce vorbesc și ce transmit? Cine interacționează cu conținutul lor? Ce efect ar putea avea acest conținut asupra cititorilor?

STABILIȚI CREDIBILITATEA: Verificați în mod independent informațiile cu surse credibile, cum ar fi experți cu experiență practică în domeniu. Ceea ce este descris are sens pe baza cunoștințelor dumneavoastră?

Folosiți tehnicile **FORENSIC MEDIA** pentru a completa reportajul de investigație tradițional și cercetarea documentară. Unele tehnici includ: extragerea de date, geolocalizarea, recunoașterea biometrică, analiza modelelor și multe altele.

ÎNVĂȚAȚI ȘI ADAPTAȚI-VĂ: Creatorii de dezinformare generată de AI se adaptează în mod constant. Adaptați-vă abordarea la peisajul în schimbare.

ÎMPĂRTĂȘEȘTE-ȚI MUNCA

Alături de o afirmație demontată, experții recomandă furnizarea unei analize transparente și a unor linkuri către surse. Acest lucru îi poate ajuta pe cititori să urmărească o investigație și să înțeleagă o narațiune nuanțată. În unele cazuri, investigația este mai importantă decât dacă conținutul este redactat de AI.

*** Henk van Ess este expert în OSINT și în tehnici de fact-checking.*

EFCSN

Ghid rapid de referință: Ce trebuie făcut, ce nu trebuie făcut și indicii

Următoarele sunt indicii care ar putea arăta că un conținut este generat de inteligența artificială sau alterat digital. Împreună cu celelalte sfaturi menționate în acest ghid (context, tehnici de investigare și instrumente de detectare), acestea vă pot ajuta să înțelegeți adevărul din spatele a ceea ce vedeți.

Text

- Adesea (dar nu întotdeauna) are o **gramatică mai bună** decât un om.
- Este probabil să folosească un **limbaj prea formal sau structurat**, în special pentru un context social media.
- **Exces de adverbe sau adjective.**
- Lipsa emoțiilor umane, a umorului, a sarcasmului și a expresiilor idiomatice.
- Poate fi **lipsit de detalii specifice** (nume, date, locații) sau de idei originale.
- Cel mai important: sunt corecte faptele afirmate în text?

Audio

- **Comparați înregistrările audio suspecte cu o mostră autentică,** utilizând instrumente care pot detecta diferențe în tiparele de vorbire și respirație, intonație...
- Atunci când utilizați detectoare, evitați eșantioanele audio de calitate scăzută cu zgomot static sau de fond.
- Poate fi marcată de **modele de vorbire nefirești sau mecanice**, lipsa pauzelor sau a respirației naturale.

Video

- Nu utilizați un detector pentru imagini generate de AI pe capturi dintr-un videoclip.
- Observați **expresiile faciale și mișcările**, cum ar fi clipitul, și dacă mișcarea gurii se potrivește cu cea a sunetului.
- Poate fi marcată de tranziții sau **tăieturi bruște.**

Imagini

- Căutați zonele cu **detalii nefirești**: piele perfectă, fundaluri neclare, frumusețe sau lumină nefirească și ciudățenii, cum ar fi degetele adiționale.
- **Căutați un filigran** al generatoarelor obișnuite de imagini.
- Fiți atenți la detaliile a ceea ce este descris: este logic? Este adecvat?
- Atunci când utilizați detectoare, **optați pentru o versiune de înaltă rezoluție sau pentru o versiune inițială în locul uneia care a fost distribuită și redistribuită.**