EFCSN

Com detectar i entendre contingut generat per IA o alterat digitalment

Els models d'IA accessibles i fàcils d'utilitzar poden ajudar a aprendre i crear contingut, però també **a amplificar els riscos que la desinformació suposa** per a les societats obertes i el discurs democràtic. És important evitar que els espais d'informació compartits s'omplin de desinformació creada amb IA o continguts manipulats digitalment.

Part de la solució són les noves tecnologies, com ara iniciatives de rastreig de la procedència dels continguts i programes de detecció. **Però les solucions tecnològiques no són perfectes i necessitem el treball d'organitzacions de verificació independents per oferir a la societat un conjunt compartit de dades verificades.**

Aquí tens una visió general ràpida del que fan els *fact-checkers* professionals i independents per identificar i desmentir la desinformació generada per IA i què es pot aprendre del seu mètode.

El contingut generat amb IA va en augment

Actualment, desinformació generada amb IA és només una petita part de les afirmacions a verificar per *fact-checkers* professionals i independents. El contingut manipulat amb eines digitals ocupa una bona part de les tasques de verificació.

Però, en una enquesta interna feta als membres de EFCSN, **la majoria dels verificadors coincideixen que el contingut generat amb IA i la manipulació digital seran més rellevants en el futur**. Exemples recents en el context de les Eleccions Europees recolzen aquesta prognosi.

CONCEPTE: ens referim a *manipulació digital* per descriure qualsevol mena de contingut que ha estat alterat significativament per tal de manipular missatge transmès originalment, això inclou eines d'IA. No inclou edicions per claredat i qualitat.

Generat amb IA: ens referim a qualsevol contingut creat per un sistema d'intel·ligència artificial.

Informació proporcionada per **AI@EUElections**, fundat per Meta i gestionat per l'**European Fact-checking Standards Network** en associació amb els seus membres i experts externs.

La tecnologia és ràpida, però no podem dependre únicament d'ella

Els experts en IA i els verificadors professionals estan d'acord: **les eines de detecció d'IA no són suficients per detectar el contingut generat per IA o alterat digitalment.**

Abans d'utilitzar un detector, els experts recomanen familiaritzar-se amb els generadors i detectors de contingut d'IA. Comprendre com s'entrenen els models i saber una mica d'estadístiques és el que necessiten els verificadors per reconèixer els punts forts i els febles d'una eina i la seva probabilitat d'èxit. **Tanmateix, les eines poden ser un bon punt de partida.**

Procedència del contingut: iniciatives com les especificacions C2PA poden ajudar a certificar l'origen i l'historial del contingut, però les marques d'aigua i les verificacions no són perfectes.

Desinformació: com ens afecta

"Cada cop que reaccionem, si es pot dir, des de les entranyes, ens saltem la reflexió"
- Christine Dugoin*

PSICOLOGIA: Influencia operacions que sovint estan dissenyades per aprofitar-se dels biaixos psicològics.

Entendre els nostres i els de la nostra audiència ens pot ajudar a combatre la desinformació.

OBJECTIUS: Per què un actor malintencionat pot fer servir IA per crear o compartir desinformació? Quin impacte en el món està buscant?

- ➔ Expandir el seu abast a una altra comunitat o territori?
- ➔ Evitar ser detectat pels verificadors o saturar-los amb moltes variants d'una mateixa afirmació?
- ➔ Establir credibilitat amb xarxes de comptes falsos per influenciar pensaments o opinions?

** Christine Dugoin és una investigadora d'influència informacional a La Sorbonne.*

Desmentir requereix un enfocament polifacètic i una comprensió matisada

Aleshores, si les eines de detecció no funcionen, què ho fa? Cal entendre tant el context d'una reclamació com el seu contingut. Els verificadors de fets professionals són experts en les habilitats d'investigació necessàries. Aquí teniu alguns consells.

> *"Les eines de detecció mai seran 100% efectives —i no espero que ho siguin."*
> – Henk van Ess**

CONSIDERA LA FONT: Pots confirmar la identitat de la font? De què parla i què comparteix? Qui interactua amb el contingut? Quin efecte pot tenir en els lectors?

ESTABLEIX CREDIBILITAT: Verifica la informació de manera independent amb fonts creïbles, com ara experts amb experiència pràctica en el camp. Allò que es representa té sentit segons els teus coneixements?

Usa tècniques d'anàlisi **FORENSE DIGITAL** a més de la investigació tradicional i la cerca documental. Algunes són: *scraping* de dades, geolocalització, reconeixement biomètric, anàlisi de patrons i moltes més.

APRÈN I ADAPTA'T: Els creadors de desinformació generada per IA s'adapten constantment. Ajusta la teva estratègia a aquest terreny en constant transformació.

COMPARTEIX LA TEVA FEINA

Juntament amb l'afirmació desmentida, els experts recomanen proporcionar una anàlisi transparent i enllaçar fonts. Això pot ajudar els lectors a seguir una investigació i entendre una narrativa matisada. En alguns casos, la investigació és més important que el fet que el contingut estigui escrit amb IA.

*** Henk van Ess és un expert en OSINT i tècniques de verificació.*

Guia ràpida de referència: Què cal fer i pistes

A continuació es mostren pistes que poden indicar que un contingut està generat per IA o alterat digitalment. Juntament amb els altres consells esmentats en aquesta guia (context, tècniques d'investigació i eines de detecció), us poden ajudar a entendre la veritat darrere del que veieu.

Text

- Sovint (però no sempre) té **millor ortografia i gramàtica** que un humà.
- És probable que utilitzi **un llenguatge massa formal**, especialment pel context de les xarxes socials.
- **Ús excessiu d'adverbis o adjectius.**
- Manca d'emoció, humor, sarcasme i expressions habituals
- Li poden **mancar detalls específics** (noms, dates, llocs) o idees originals.
- El més important: són correctes les dades que afirma el text?

Àudio

- **Compara clips d'àudio sospitosos amb la mostra original fent servir eines que detectin diferències en la parla, en els patrons de respiració, en l'entonació...**
- Quan utilitzis detectors, evita mostres d'àudio de baixa qualitat amb estàtica o soroll de fons.
- Pot estar marcat **per patrons de parla no naturals o mecànics**, manca de pauses o respiració natural.

Vídeo

- No facis servir eines de detecció d'IA en una captura estàtica del vídeo.
- Mira **les expressions facials i el moviment** com ara el parpelleig. També si el moviment de la boca coincideix amb l'àudio.
- Es pot marcar amb **transicions o talls nítids**.

Imatges

- Busca àrees amb detalls **poc naturals**: pell perfecta, fons borrós, llum no natural i defectes com una mà amb dits de més.
- **Busca la marca d'aigua** de generadors d'imatges habituals.
- Fixa't en els detalls: el que es representa té sentit? És apropiat?
- Detectors: **opta per una versió d'alta resolució o una versió primerenca de la imatge en lloc d'una que s'hagi compartit una vegada i una altra.**