EFCSN

Come riconoscere i contenuti generati dall' IA o alterati digitalmente

Modelli di Intelligenza Artificiale (IA) accessibili e facili da usare possono aiutare le persone ad imparare e a creare contenuti, ma possono anche **amplificare i rischi che la mis- e disinformazione comportano** per le società aperte e il sistema democratico. Bisogna impedire che gli spazi di informazione si riempiano di contenuti disinformativi generati dall'IA e alterati digitalmente.

Le nuove tecnologie, come i software di rilevamento della provenienza dei contenuti, possono essere parte della soluzione ma sono lungi dall'essere perfette. **Abbiamo bisogno del lavoro di fact-checker indipendenti per fornire alla società una serie condivisa di fatti verificati.**

Ecco una rapida panoramica di ciò che fanno i fact-checker professionisti e indipendenti per verificare i contenuti mis- e disinformativi generati dall'IA e cosa puoi imparare da loro.

Aumento dei contenuti generati dall'IA

Oggi, la mis- e disinformazione generata dall'IA rappresenta una piccola percentuale di tutte le affermazioni esaminate da fact-checker professionisti e indipendenti. I contenuti alterati digitalmente sono diffusi soprattutto nell'ambito di lavoro specifico dei fact-checker.

Ma in un sondaggio interno condotto tra i membri dell'EFCSN è emerso che la **maggior parte dei fact-checker concorda che i contenuti generati dall'intelligenza artificiale e alterati digitalmente non potranno che aumentare** di rilevanza in futuro. Esempi recenti nel contesto delle elezioni europee sembrano supportano questa previsione.

CONCETTI CHIAVE: *Digitalmente alterato* si riferisce a qualsiasi forma di contenuto che è stato alterato in modo significativo per manipolare o cambiare il messaggio originariamente trasmesso, comprese le modifiche apportate dagli strumenti di intelligenza artificiale. Ciò non include le modifiche per migliorare chiarezza o qualità.

Generato dall'IA si riferisce a qualsiasi forma di contenuto creato da un sistema di intelligenza artificiale.

Informazioni fornite da **AI@EUElections**, un progetto finanziato da Meta e gestito dallo European Fact-Checking Standards Network in associazione con i suoi membri ed esperti esterni.

La tecnologia avanza rapidamente, ma non possiamo farci affidamento esclusivo

Gli esperti di IA e i fact-checker concordano: **gli strumenti di rilevamento dell'intelligenza artificiale da soli non sono sufficienti per rilevare o verificare i contenuti generati dall'intelligenza artificiale o alterati digitalmente**.

Prima di utilizzare un rilevatore, gli esperti consigliano di acquisire familiarità con i generatori e i rilevatori di contenuti IA. Capendo come vengono addestrati i modelli, e con qualche nozione di statistica, i fact-checker possono iniziare a riconoscere i punti di forza e di debolezza di uno strumento e la sua probabilità di successo. In ogni caso, **gli strumenti tecnologici possono essere un utile punto di partenza.**

Gli strumenti tecnologici che rilevano la **provenienza dei contenuti,** come il protocollo C2PA , possono aiutare a certificare la fonte e la storia dei contenuti multimediali, ma la verifica non è affidabile al 100%

Come la disinformazione basata sull'IA impatta le persone

"Ogni volta che reagisci di pancia bypassi la riflessione"
– Christine Dugoin*

PSICOLOGIA: Le operazioni volte ad influenzare l'opinione pubblica sono spesso progettate per trarre vantaggio dai pregiudizi psicologici. Capire quali sono i tuoi e quelli della tua audience può aiutare a contrastare la disinformazione.

OBIETTIVI: Perché un malintenzionato potrebbe fare affidamento sull'IA per creare o diffondere disinformazione? Qual è il l'impatto sperato nel mondo reale?

- ➔ Espandere la loro portata in un altro paese o comunità?
- ➔ Evitare di essere scoperti o confondere i fact-checker generando molte varianti di affermazioni simili?
- ➔ Influenzare pensieri o convinzioni creando credibilità attraverso reti di account fittizzi?

** Christine Dugoin è una ricercatrice in influenza dei media a La Sorbona.*

Il debunking richiede un approccio integrato

Quindi, se gli strumenti di rilevamento non funzionano, cosa funziona? È importante comprendere il contesto di una affermazione tanto quanto il suo contenuto. I fact-checker professionisti possiedono le competenze investigative necessarie. Ecco alcuni suggerimenti.

> *"Gli strumenti di rilevamento automatico non funzioneranno mai al 100% - non prevedo che lo faranno mai."*
> - Henk van Ess**

ESAMINA LA FONTE: Puoi confermarne l'identità? Di cosa parlano e cosa condividono? Chi interagisce con i loro contenuti? Che effetto potrebbe avere questo contenuto sui lettori?

VALUTA LA CREDIBILITA': Verifica le informazioni attraverso fonti affidabili come esperti con esperienza pratica sul campo. Ciò che viene sostenuto ha senso in base alle tue conoscenze?

Utilizza le tecniche di **MEDIA FORENSIC** per integrare il tradizionale reporting investigativo e la ricerca documentaria. Alcune tecniche includono: raccolta di dati, geolocalizzazione, riconoscimento biometrico, analisi di modelli e altro ancora.

IMPARA & ADATTATI: i creatori di contenuti di mis- e disinformazione generati dall'IA si adattano costantemente. Adatta anche tu il tuo approccio al panorama in evoluzione.

CONDIVIDI IL TUO LAVORO

Gli esperti raccomandano di fornire, insieme al contenuto verificato, anche i vari step dell'analisi che si è svolta e i collegamenti alle fonti. Ciò può aiutare i lettori a comprendere meglio il processo di verifica e la narrazione.

In alcuni casi, il processo di indagine stesso è ancora più importante del suo esito riguardo a se il contenuto sia stato effettivamente generato dall'IA oppure no.

*** Henk van Ess è un esperto di OSINT e tecniche di fact-checking.*

Guida essenziale: Cosa fare, non fare e indizi

Di seguito sono riportati alcuni indizi che potrebbero indicare che un contenuto è stato generato dall'IA o alterato digitalmente. Insieme agli altri suggerimenti citati in questa guida (contesto, tecniche di indagine e strumenti di rilevamento), possono aiutarti a comprendere la verità dietro a ciò che vedi.

Testo

- Spesso (ma non sempre) ha una **grammatica migliore** di un essere umano.
- È probabile che utilizzi un **linguaggio eccessivamente formale** o strutturato, specialmente per il contesto dei social media.
- **Avverbi o aggettivi eccessivi.**
- Mancanza di emozioni umane, umorismo, sarcasmo ed espressioni idiomatiche.
- Potrebbero **mancare dettagli specifici** (nomi, date, luoghi) o idee originali.
- Cosa più importante: i fatti riportati nel testo sono corretti?

Video

- Non utilizzare un rilevatore di immagini generato dall'IA su immagini fisse di un video.
- Osserva le **espressioni facciali e i movimenti**, come il battito delle palpebre, e se il movimento della bocca corrisponde all'audio.
- Può essere caratterizzato da **transizioni o tagli netti**.

Audio

- **Confronta l'audio sospetto con un campione autentico** utilizzando strumenti in grado di rilevare differenze nei modelli di linguaggio e respirazione, nell'intonazione...
- Quando usi i rilevatori, evita campioni audio di bassa qualità con rumore statico o di sottofondo.
- Può essere caratterizzato da **schemi di linguaggio innaturali o meccanici**, mancanza di pause o respiro naturale.

Immagini

- Cerca zone con **dettagli innaturali**: pelle perfetta, sfondi sfocati, bellezza o luce innaturali e stranezze come dita in più.
- **Cerca un' eventuale watermark** che spesso generatori di immagini comuni lasciano sull'immagine.
- Presta attenzione ai dettagli di ciò che è raffigurato: è logico? È appropriato?
- Se si usi i rilevatori, **scegli una versione ad alta risoluzione dell'immagine o una delle prime versioni caricate**, anziché una che è già stata condivisa e ricondivisa.