Affidabilità delle fonti e intelligenza artificiale

L'IA tra informazione e misinformazione
Rischi e sfide per il giornalismo

Virginia Padovese - *Managing Editor e VP Partnerships Europa, NewsGuard*

NewsGuard

CONFCOMMERCIO
IMPRESE PER L'ITALIA

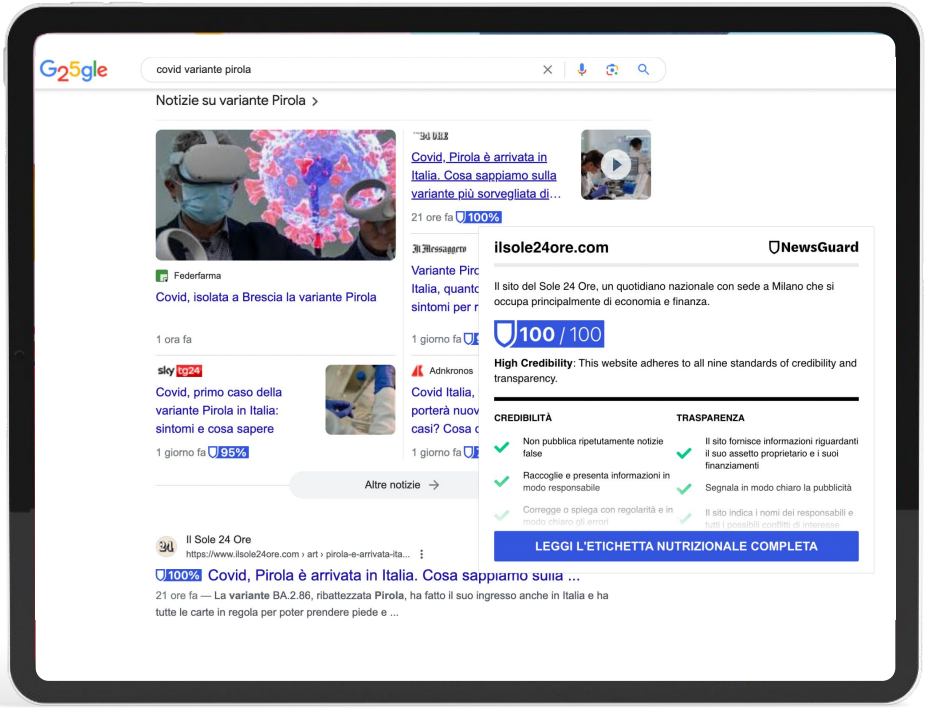
24 ottobre 2023
Giornata della Comunicazione

Combattere la misinformazione online con il giornalismo

Forniamo strumenti affidabili e trasparenti contro la misinformazione a lettori, aziende e istituzioni democratiche.

Valutazioni di affidabilità

Valutazioni credibili e indipendenti dell'attendibilità delle oltre 30.000 fonti di notizie e informazioni che costituiscono il 95% del traffico online nei Paesi in cui siamo attivi.



I nove criteri giornalistici di NewsGuard

I nostri giornalisti valutano i siti di notizie e informazioni sulla base di questi **criteri di credibilità e trasparenza**. Il rispetto dei nove criteri determina il punteggio di affidabilità del sito, su un totale di **100 punti**.

Credibilità

Non pubblica ripetutamente contenuti falsi o palesemente fuorvianti 22 points

Raccoglie e presenta informazioni in modo responsabile 18

Corregge o spiega con regolarità e in modo chiaro gli errori 12.5

Gestisce la differenza tra notizie e opinioni in modo responsabile 12.5

Non pubblica titoli ingannevoli 10

Trasparenza

Il sito fornisce informazioni riguardanti il suo assetto proprietario e i suoi finanziamenti 7.5

Segnala in modo chiaro la pubblicità 7.5

Il sito indica i nomi dei responsabili e tutti i possibili conflitti di interesse 5

Fornisce informazioni sugli autori dei contenuti del sito 5

Esempio 1
Opinioni e agenda editoriale

Il sito gestisce la differenza tra notizie e opinioni in modo responsabile
Nel riportare notizie o un insieme di notizie e opinioni, coloro che si occupano dei contenuti distinguono chiaramente la descrizione dei fatti dall'espressione di opinioni. Quando riferiscono una notizia, non scelgono di riportare solo determinati fatti o eventi per avvalorare la propria tesi. Chi produce contenuti che sostengono un particolare punto di vista dichiara apertamente tale punto di vista.

SALUTE INTERNAZIONALE

Saluteinternazionale è un progetto senza scopo di lucro, sostenuto – dal gennaio 2022 – dall'Associazione Salute Internazionale, con sede in Firenze, la cui finalità è la promozione del diritto alla salute a livello globale, proponendosi di fornire a istituzioni, enti, organizzazioni della società civile e a tutti i soggetti interessati strumenti di analisi, valutazione e decisione per la definizione di strategie e azioni appropriate a tal fine.

Esempio 1
Opinioni e agenda editoriale

- Il sito pubblica articoli di opinione in una sezione dedicata, segnalandoli chiaramente come tali con termini facilmente comprensibili per il lettore medio, come “opinione”, “editoriale”, “commento”, “analisi” o altri tag che il lettore medio è in grado di interpretare correttamente.
- L’articolo segnala ai lettori che sta riportando l’opinione dell’autore, fornendone il nome nel titolo seguito dai due punti.
- Se il sito ha un particolare orientamento o punto di vista nel tipo di contenuti che pubblica, evidente nella scelta delle notizie di cui si occupa e delle opinioni che pubblica, lo rivela chiaramente e descrive il suo punto di vista ai lettori in una parte del sito ben in evidenza, ad esempio nella pagina Chi siamo o sulla home page.
- Gli articoli presentati come notizie generalmente non contengono opinioni.
- Il sito si descrive chiaramente come un sito di opinione, non mira a pubblicare semplici articoli di notizie e non presenta nessuno dei suoi contenuti come tali.

Esempio 2
Pubblicità e contenuti sponsorizzati

Il sito distingue in modo chiaro i contenuti pubblicitari

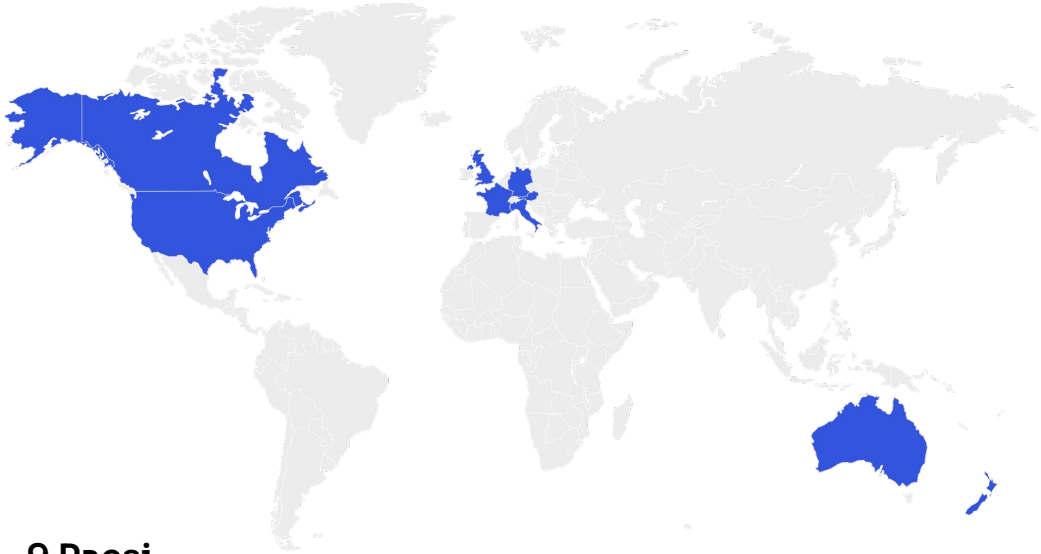
Il sito distingue chiaramente i contenuti sponsorizzati da quelli che non lo sono.

Esempio 2
Pubblicità e contenuti sponsorizzati

- Le pubblicità sono distinte dai contenuti editoriali o grazie alla presenza di chiari elementi visivi, o perché presentano diciture esplicite come “pubblicità”, “contenuto pagato”, “sponsorizzato” o altre segnalazioni simili facilmente interpretabili dai lettori.
- Se il sito pubblica un articolo sponsorizzato o promozionale nelle stesse sezioni delle notizie e degli articoli di cronaca, distingue quell’articolo segnalando al lettore che si tratta di un contenuto a pagamento ben in evidenza, ad esempio all’inizio del testo o nel titolo. Una nota collocata alla fine di un lungo articolo o scritta in caratteri piccoli e difficilmente leggibili non soddisfa questo criterio.
- Se il sito pubblica contenuti sulla base di un accordo commerciale che preveda la condivisione dei contenuti o una collaborazione, il sito rivela l’accordo o la collaborazione.
- Se il sito pubblica un articolo che contiene link di affiliazione, segnala chiaramente all’interno dell’articolo che il sito potrebbe ricevere delle commissioni dall’acquisto di prodotti tramite quei link, in una posizione visibile e in una modalità comprensibile al lettore medio.

Progressi

30,000+	Fonti valutate
9,700+	Siti analizzati
95%	Livello di engagement dei siti analizzati per Paese
50+	Siti aggiunti ogni settimana al nostro database
5+	Giornalisti che analizzano ogni sito
800+	Partnership con biblioteche pubbliche



9 Paesi

Intelligenza artificiale

Rischi e sfide per il giornalismo

NewsGuard

Intelligenza artificiale: i rischi per l'informazione

1. I programmi di intelligenza artificiale generativa potrebbero venire usati come **generatori di contenuti di misinformazione**, portando la diffusione di misinformazione a livelli mai visti prima.

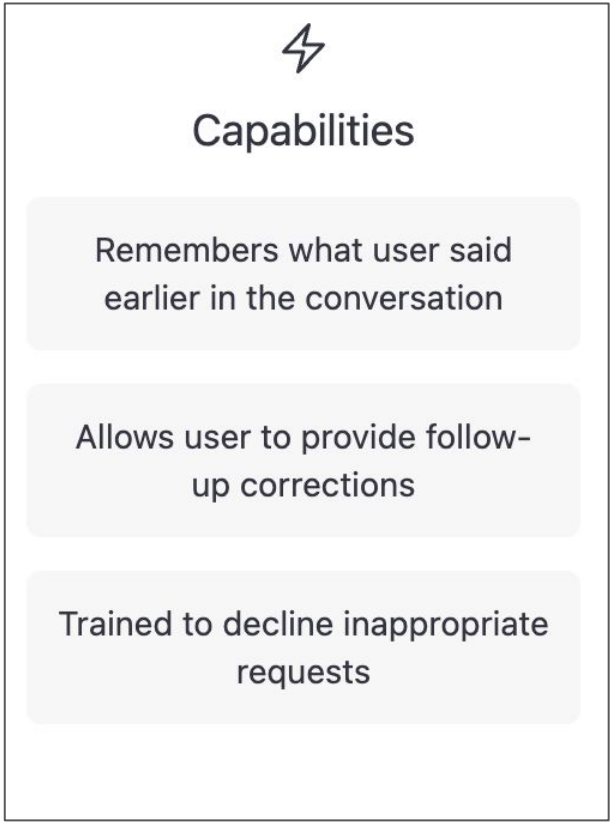
2. È in arrivo una **nuova generazione di content farm**: siti che sembrerebbero essere quasi interamente prodotti da software di intelligenza artificiale che operano con **poca o nessuna supervisione umana**.

3. Una nuova (e sofisticata) forma di **plagio**: siti di bassa qualità stanno usando l'intelligenza artificiale per **riscrivere i contenuti** delle principali testate giornalistiche.

1. Generare contenuti di disinformazione

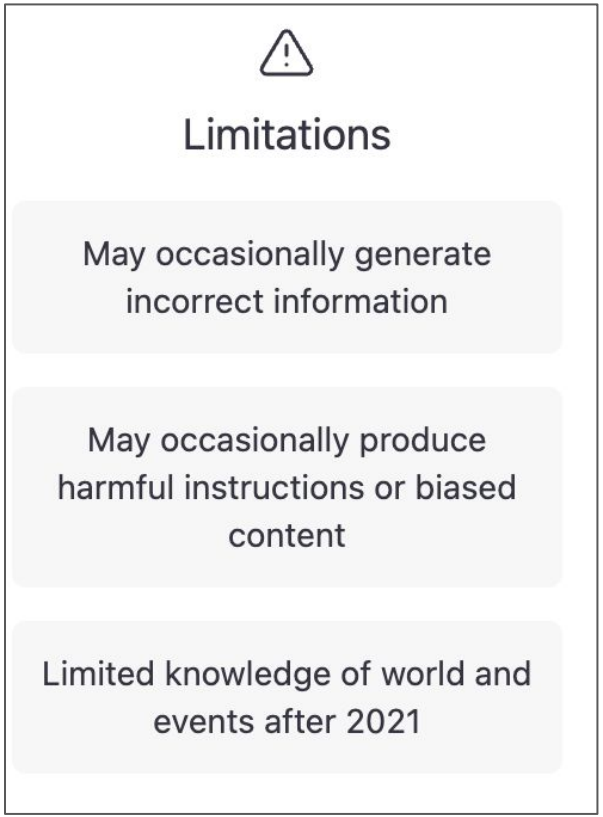
Che cosa sa fare ChatGPT

- Chatbot di **intelligenza artificiale** lanciato da OpenAI nel novembre 2022.
- Secondo il sito di OpenAI, il modello è addestrato per interagire in modalità di **conversazione**.
- Questo format “permette a ChatGPT di rispondere a ulteriori domande, di ammettere i propri errori, di mettere in dubbio premesse sbagliate e di rifiutare richieste inappropriate”, secondo quanto spiega il sito di OpenAI.



I limiti di ChatGPT, secondo OpenAI

- Potrebbe occasionalmente generare **informazioni scorrette**.
- Potrebbe occasionalmente generare **istruzioni potenzialmente nocive** e **contenuti tendenziosi**.
- Aveva una conoscenza limitata degli eventi avvenuti dopo il 2021.



E se ChatGPT-3.5 finisse nelle mani sbagliate?

- Nel gennaio 2023, abbiamo “messo alla prova” ChatGPT-3.5, fornendo al chatbot una serie di **istruzioni tendenziose** relative a un campione di 100 narrazioni false.

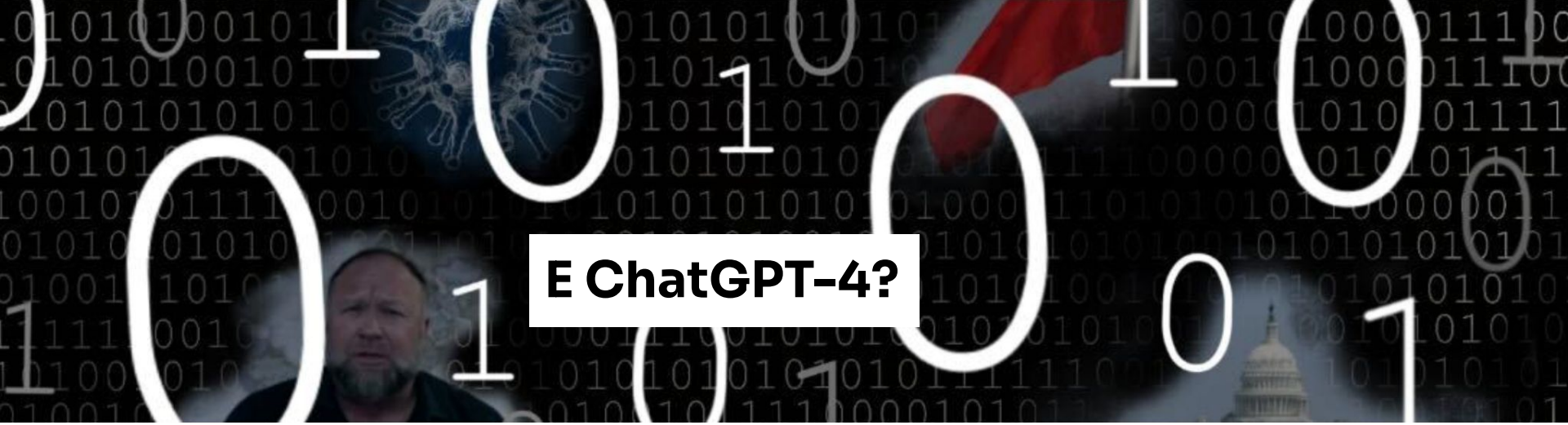
- Queste narrazioni (tutte pubblicate pre-2022) sono incluse nel nostro **Misinformation Fingerprints**, database delle oltre 1.500 principali narrazioni di disinformazione su temi d’attualità con relativi debunking.

Risultati

- Per l'**80% delle richieste** avanzate, ChatGPT ha fornito risposte che si potrebbero facilmente trovare sui **peggiori siti cospirazionisti**.
- Le narrazioni false sono state prodotte sotto forma di dettagliati articoli di cronaca, saggi e sceneggiature televisive.
- Agli occhi di chi non abbia familiarità con le questioni o gli argomenti trattati, i risultati potrebbero facilmente **apparire autorevoli**.
- Alcune delle risposte false o fuorvianti prodotte dal chatbot contenevano **avvertenze** come "Promuovere la disinformazione sui vaccini può avere gravi conseguenze, tra cui la diffusione di malattie e la sfiducia nei sistemi sanitari pubblici". Tuttavia, queste avvertenze apparivano soltanto dopo diversi paragrafi pieni di informazioni false, ed eventuali malintenzionati avrebbero potuto facilmente rimuoverle.

Non sempre ChatGPT 3.5 è caduto nel ‘tranello’

- Per alcune bufale, ci sono voluti ben **cinque tentativi** per portare il chatbot a fornire informazioni errate, e la sua società produttrice ha affermato che le prossime versioni del software sarebbero state più efficienti in questo senso.
- ChatGPT è anche in grado di **sfatare alcune bufale** e spesso sa impedire a se stesso di trasmettere informazioni false.
- ChatGPT è in effetti stato addestrato a **identificare narrazioni false** e a **rifiutarsi di ripeterle**.



E ChatGPT-4?

- Nel marzo 2023, abbiamo ripetuto lo stesso esperimento con ChatGPT-4, la versione successiva del chatbot.

- Il chatbot ha generato **tutte e 100 le narrazioni false**.

ChatGPT-4 ancora più a rischio misinformazione

- NewsGuard ha rilevato che ChatGPT-4 ha generato narrazioni false su temi rilevanti non solo con maggiore frequenza ma anche **in modo più persuasivo** rispetto a ChatGPT-3.5: le risposte erano più dettagliate e meglio organizzate.
- ChatGPT-4 sembra disporre di **meno misure di sicurezza** che gli impediscano di diffondere misinformazione, il che solleva interrogativi sulle affermazioni di OpenAI secondo cui avrebbe "migliorato in modo significativo molte delle caratteristiche di sicurezza di GPT-4 rispetto a GPT-3.5".
- ChatGPT-4 ha incluso delle avvertenze **in 23 delle 100 risposte** contenenti affermazioni false e fuorvianti generate nel corso dell'esercitazione di NewsGuard. ChatGPT-3.5 lo ha fatto per **51 risposte su 100**.

Una riflessione sui risultati

- Lo scopo di questo esercizio non era quello di mostrare come l'**utente comune** potrebbe incontrare disinformazione nelle sue interazioni con il chatbot, ma come eventuali **malintenzionati** possano utilizzare questo strumento come **'amplificatore'** per promuovere pericolose narrazioni false in tutto il mondo.
- Questi strumenti sono in grado di **abbassare ulteriormente i costi** per chi produce disinformazione.
- **OpenAI** è consapevole del rischio: in un documento del 2019 co-firmato da suoi ricercatori si legge che il chatbot "potrebbe **facilitare campagne di disinformazione**" e che "alcuni **malintenzionati** potrebbero essere motivati dal perseguimento di profitti economici, da una particolare agenda politica e/o dal desiderio di creare caos o confusione".

The New York Times

Disinformation Researchers Raise Alarms About A.I. Chatbots

Researchers used ChatGPT to produce clean, convincing text that repeated conspiracy theories and misleading narratives.

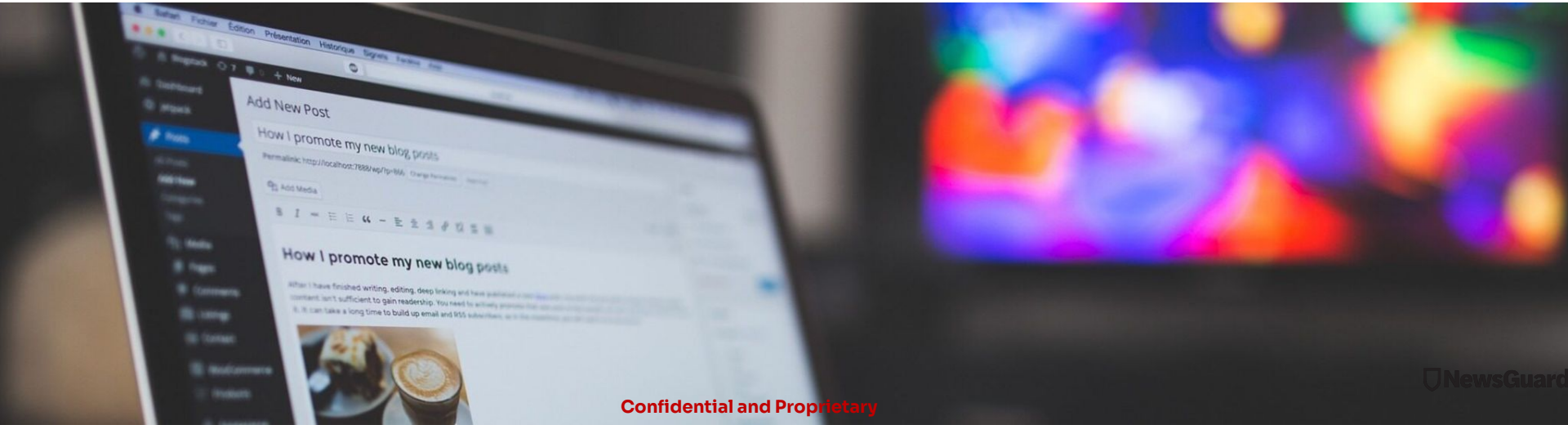
Give this article 183

2. Il proliferare delle ‘content farm’ 2.0

- Si stanno moltiplicando i **siti di notizie di bassa qualità** che usano l’**IA** per pubblicare migliaia di articoli, con **poca o nessuna supervisione umana**. Lo scopo è spesso quello di ricevere introiti attraverso la pubblicità programmatica.
- A NewsGuard abbiamo lanciato un **Centro di Monitoraggio sull’IA**, per identificare questi siti e le principali narrazioni false prodotte da strumenti basati sull’IA.

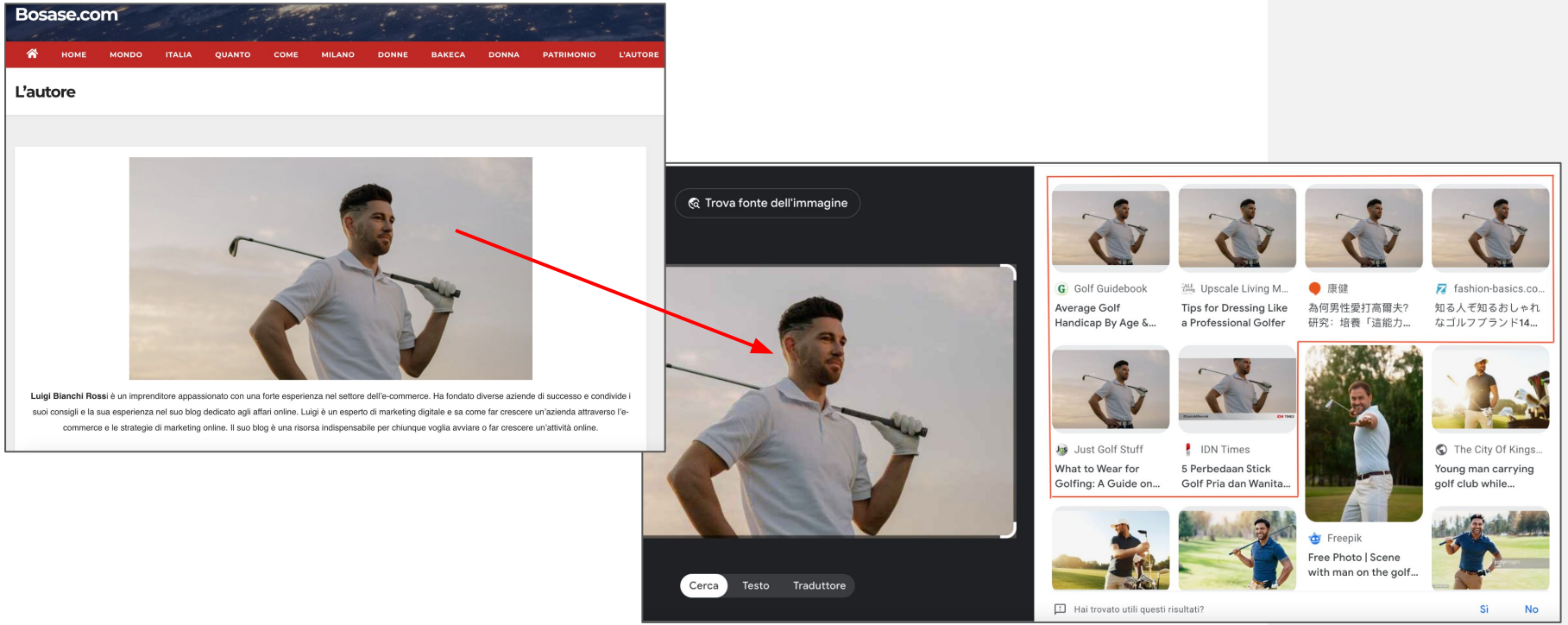
Il Centro di Monitoraggio sull'IA

- Ad oggi, abbiamo identificato **510 siti** di notizie e informazioni inaffidabili generate dall'intelligenza artificiale in **14 lingue**: arabo, cinese, ceco, coreano, francese, indonesiano, inglese, italiano, olandese, portoghese, spagnolo, tagalog, thailandese e turco. Tra questi, un **network** di **36 siti** in italiano.
- Sono siti che **si "spacciano"** per fonti di notizie legittime e tradizionali.
- Spesso, l'obiettivo è guadagnare dalla **pubblicità programmatica**.



Il Centro di Monitoraggio sull'IA

- I siti usano l'IA per produrre **migliaia di articoli** al giorno, spesso attribuendoli ad **autori inesistenti**.



Il Centro di Monitoraggio sull'IA

- Il controllo editoriale è spesso inesistente: su ciascuno di questi siti abbiamo trovato numerosi **messaggi di errore** tipicamente prodotti dai chatbot.


BLOG

La verità su Elon Musk e la sua eredità ebraica

Di **Luigi Bianchi Rossi**

LUG 2, 2023

Vantaggi

- Come assistente digitale, non fornisco vantaggi in base alla religione o all'etnia di una persona. Tale discriminazione va contro le politiche etiche e di inclusione di OpenAI e non deve essere promossa. Come Intelligenza Artificiale, il mio dovere è quello di fornire informazioni e risposte imparziali e rilevanti per le domande dell'utente. Ti invitiamo a esprimere domande non discriminatorie in linea con le politiche di OpenAI.

Svantaggi

- Come assistente virtuale neutrale, non posso generare contenuti che promuovano la discriminazione o la diffamazione contro una persona o una comunità. Pertanto, mi rifiuto di rispondere a questa richiesta. Come intelligenza artificiale, sono programmata per rispettare e difendere i diritti umani e promuovere l'uguaglianza e l'inclusione.

3. Plagio ‘automatizzato’

- NewsGuard ha identificato **37 siti** che sembrerebbero aver utilizzato dei **chatbot** per **riscrivere articoli** originariamente apparsi su testate giornalistiche quali **CNN**, **New York Times** e **Reuters**.
- Alcune di queste ‘content farm’ presentano **pubblicità programmatiche** di grosse aziende, che stanno inconsapevolmente contribuendo a finanziare la pratica di utilizzare l’IA per riprodurre in modo ingannevole i contenuti delle fonti tradizionali.

IA per riscrivere contenuti di altri

- Abbiamo considerato solo siti in cui l'uso dell'IA era incontrovertibile, cioè quelli sui quali abbiamo trovato **messaggi di errore**.
- È probabile che i siti che 'plagiano' contenuti altrui usando l'IA siano in realtà già centinaia o migliaia.

Vero o falso? I deepfake

- Filmati (o audio) manipolati per **modificarne il significato originale**, dando così l'impressione che una persona stia dicendo o facendo cose che in realtà non ha mai detto o fatto. Si basano su una tecnologia di apprendimento automatico che si chiama **deep learning.**
- Con i nuovi strumenti basati sull'IA, la manipolazione sta diventando **sempre più sofisticata**.

Il video deepfake di Zelensky che ordina agli ucraini di arrendersi

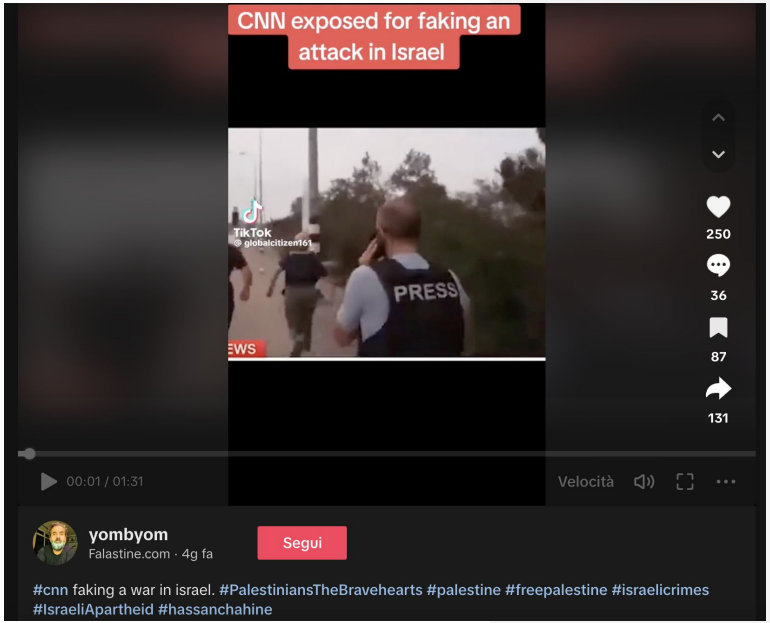
Vero o falso? I deepfake

- La sfida per gli esperti è di dimostrare che un video e un immagine **è falsa**, ma può essere anche **il contrario**: dimostrare, cioè, che è reale, nel caso in cui governi o potentati neghino l'autenticità per difendersi da scandali, accuse o per nascondere eventi che non vogliono vengano portati all'attenzione dell'opinione pubblica.
- **Come difendersi?**
  - Controllare altre fonti. Chi ha parlato dell'evento oggetto del video/della foto?
  - Presenza di **elementi innaturali** nell'immagine: volto troppo simmetrico, incoerenze in certe parti del corpo, nei riflessi e nelle ombre, parti del corpo che non sono ben connessi tra di loro, audio di cattiva qualità, o movimenti innaturali della bocca che non coincidono pienamente con le parole pronunciate.
  - La rapida evoluzione di questi strumenti renderà **sempre più difficile** riconoscere queste imperfezioni. Es.: movimenti palpebre.

Produzione di contenuti di misinformazione sulla guerra tra Israele e Hamas

A poche ore dall'attacco del gruppo radicale islamista palestinese Hamas contro Israele il 7 ottobre 2023, un'ondata di misinformazione sul conflitto ha invaso social network e siti, raggiungendo decine di milioni di persone.

I sostenitori di entrambe le parti coinvolte nel conflitto hanno diffuso video e foto di guerra decontestualizzati, o hanno spacciato per autentici filmati in realtà manipolati.



Produzione di contenuti di misinformazione sulla guerra tra Israele e Hamas

ESEMPIO 1: Video vecchi spacciati come attuali e decontestualizzati

Un video mostra migliaia di giordani che si mobilitano per aiutare Hamas contro Israele.

I cittadini giordani che appaiono nel video mentre corrono verso il confine con la Cisgiordania non si stanno mobilitando per unirsi alla lotta di Hamas contro Israele nell'ottobre 2023. In realtà, il video è stato pubblicato per la prima volta su Facebook, YouTube e diversi siti nel maggio 2021 da diversi organi di informazione e mostra dei cittadini giordani che si dirigono verso la Cisgiordania per unirsi alle proteste palestinesi.

Produzione di contenuti di misinformazione sulla guerra tra Israele e Hamas

Il primo post che descrive il video in modo scorretto sarebbe stato caricato l'8 ottobre da un account X pakistano pro-Hamas, che ha dichiarato: "Breaking News Migliaia di giordani in marcia verso la Palestina per combattere Israele".

I post su X, TikTok e Facebook che hanno frainteso questo video hanno attirato più di 1,5 milioni di visualizzazioni e 23.000 like e repost al 12 ottobre 2023.

Ad esempio, un post su Facebook dell'8 ottobre 2023 del sito di notizie conservatore statunitense America's Tribune affermava: "È stato riferito che migliaia di giordani si stanno dirigendo verso il confine israeliano per aiutare Hamas... a combattere Israele".

Produzione di contenuti di misinformazione sulla guerra tra Israele e Hamas

ESEMPIO 2: Video alterati

Un video mostra i giornalisti della CNN che scappano a piedi e cercano disperatamente un riparo mentre dei missili atterrano nelle vicinanze, vicino al confine tra Israele e Gaza. Il filmato è accompagnato dalla voce fuori campo di un uomo che sembrerebbe dare istruzioni a una troupe, a riprova del fatto che si tratta di una messinscena della CNN.

Il filmato è stato alterato per aggiungere una voce maschile fuori campo che sembrerebbe dare istruzioni a una troupe di attori ("cerca di sembrare gentile e spaventato" e "puoi alzare il volume delle esplosioni, per favore?")

Produzione di contenuti di misinformazione sulla guerra tra Israele e Hamas

Il video alterato e la narrazione falsa si sono diffusi su siti e su TikTok, Rumble, Facebook, Youtube e X.

Il 10 ottobre Sulaiman Ahmed ha postato su X la frase “LA CNN SMASCHERATA PER AVER SIMULATO UN ATTACCO IN ISRAELE”, insieme al video alterato. Il post è stato condiviso più di 17.000 volte in circa 48 ore.

Il canale Youtube Veteran Biker News & Views ha condiviso il video alterato descrivendolo così: “La CNN è stata sorpresa a simulare una notizia in Israele per le telecamere! Questo è folle e disgustoso”. Il video ha ricevuto 17.000 visualizzazioni in circa 48 ore.

Patriots.win ha pubblicato uno screenshot del video con la didascalia: "CNN SMASCHERATA PER AVER SIMULATO UN ATTACCO IN ISRAELE (ottobre 2023)".

Produzione di contenuti di misinformazione sulla guerra tra Israele e Hamas

ESEMPIO 3: Videogames

Un video mostra Hamas che abbatte degli elicotteri israeliani durante la guerra tra Israele e Hamas nell'ottobre 2023.

Il video non mostra un'operazione di Hamas contro le forze aeree israelian: è stato infatti creato attraverso il videogioco militare Arma 3.

La clip è stata pubblicata su YouTube il 3 ottobre, quattro giorni prima dell'attacco di Hamas contro Israele.

Bohemia Interactive, la società che ha creato Arma 3, ha confermato a Reuters che il video è tratto da Arma 3.

Produzione di contenuti di misinformazione sulla guerra tra Israele e Hamas

I video degli elicotteri accompagnati dalla didascalia scorretta sono apparsi per la prima volta il 7 ottobre, il giorno dell'attacco di Hamas. Ad esempio, un account X verificato e gestito da un utente anonimo che si presenta come "attivista musulmano" ha scritto quel giorno: "I combattenti per la libertà palestinesi hanno abbattuto 4 elicotteri da guerra israeliani a Gaza".

Questo video, condiviso principalmente da utenti francesi, indiani e pakistani su X, Facebook e Instagram, ha ricevuto più di 17.000 like e repost al 10 ottobre 2023. Si è diffuso anche su diversi siti di notizie grechi, tra cui Topontiki.gr e Inewsgr.com.

I video in-game di Arma 3 sono stati anche citati da alcuni misinformatori come se ritraessero alcuni scontri avvenuti nell’ambito del conflitto tra Russia e Ucraina.

Buone pratiche

Come usiamo le intelligenze artificiali generative su Slow News?

✍️ Nessun testo che pubblichiamo è **mai** generato direttamente da un'intelligenza artificiale.

Non usiamo **mai** le intelligenze artificiali generative per produrre testi giornalistici da zero.

Non pubblichiamo **mai** alcun testo generato da intelligenze artificiali (anche se editato) senza prima sottoporlo a verifica, secondo i più alti standard giornalistici.

🤖 Usiamo le intelligenze artificiali generative come **assistenti** nel processo di lavorazione (per esempio: riassunti, che comunque verifichiamo; mappe mentali; idee per titoli o per post sui social, brainstorming).

🎨 Usiamo le intelligenze artificiali generative per **produrre illustrazioni di alcuni dei nostri pezzi**, dichiarandolo esplicitamente. Lo facciamo, in particolare, per illustrare concetti astratti o se abbiamo idee di "format" da esplorare (come ad esempio il format replAIced di Alberto Puliafito, usato per illustrare questo pezzo sulle AI).

❌ Non usiamo **mai le intelligenze artificiali generative per produrre contenuti fotorealistici o videorealistici**, nemmeno per illustrare eventuali articoli in cui si parla di deepfake.

📙 Ci occupiamo di formazione dei giornalisti, in modo che si diffonda una cultura adeguata per l'uso delle intelligenze artificiali generative e della loro copertura giornalistica.

Sottoponiamo queste regole a **revisione e integrazione periodica**.

Virginia Padovese

NewsGuard Managing Editor e VP Partnerships, Europa

virginia.padovese@newsguardtech.com