POTENZIALITÀ E LIMITI DELLA MODERAZIONE ALGORITMICA

L'INTELLIGENZA ARTIFICIALE CONTRIBUISCE ALLA CIRCOLAZIONE DELLE FAKE NEWS, MA PUÒ ANCHE AIUTARE A RIMUOVERLE IN AUTOMATICO. UN USO INCONTROLLATO DEI SISTEMI DI MODERAZIONE IN INTERNET METTE PERÒ A RISCHIO LA LIBERTÀ DI ESPRESSIONE, L'ACCESSO AI DATI E IL DIALOGO DEMOCRATICO. L'UE STA SVILUPPANDO UN QUADRO GIURIDICO AD HOC.

La sempre maggiore diffusione delle *fake news*, che già comporta gravi conseguenze per individui e società, richiede iniziative efficaci volte a identificarle e a prevenirne la diffusione.
I controlli manuali affidati ai cosiddetti *trusted flagger* non riescono a contenere il fenomeno. Infatti, tali controlli non sono scalabili (dato l'elevato costo e la scarsità delle competenze richieste), e riguardano il singolo messaggio e non il flusso della sua divulgazione ed elaborazione.
Infine, le decisioni degli esperti umani possono mancare di imparzialità, essendo condizionate da pregiudizi, idiosincrasie o preferenze politiche.
Pertanto, molte piattaforme ricorrono a strumenti di moderazione automatica, intesi a sostituire o più frequentemente a facilitare il lavoro dei moderatori (*figura 1*). Secondo statistiche recenti, nel secondo quadrimestre 2022 Facebook ha disabilitato più di 1,4 miliardi di falsi *account* che contribuivano alla produzione di *fake news*[1]. Mentre TripAdvisor ha rifiutato o rimosso il 3,6% delle recensioni pubblicate nel 2021, identificate come false[2].
I sistemi di moderazione delle *fake news* hanno l'obiettivo di individuare una notizia falsa e adottare una misura correttiva tesa a escluderne o limitarne la diffusione.
L'individuazione può avvenire in diversi modi:
- analisi dell'utente che diffonde la notizia o che le legge (spesso ciò avviene attraverso l'identificazione di profili *fake*)
- analisi del flusso (velocità e ampiezza) della distribuzione di una certa notizia e delle modalità con cui altri utenti interagiscono con tale flusso
- analisi del contenuto delle notizie.

Quest'ultima verifica valuta la veridicità delle principali affermazioni contenute in una notizia. A tal fine, esistono molti siti web curati da esperti che permettono il confronto diretto tra una notizia presuntivamente falsa e una notizia considerata vera.
L'analisi automatica del contenuto può avvenire anche attraverso metodi di elaborazione del linguaggio naturale (Nlp), basati sulle recenti tecniche di intelligenza artificiale, come l'apprendimento automatico (*machine learning*) e l'apprendimento profondo (*deep learning*). Queste tecniche permettono di creare un sistema che è in grado di verificare se una certa sequenza di parole o una certa struttura sintattica corrisponda statisticamente a una *fake news*.
Questi sistemi apprendono una tale correlazione direttamente dai testi oppure attraverso appositi esempi di notizie false forniti dagli esperti umani (*figura 1*).
L'individuazione automatica di una *fake news* può comportare l'adozione di diverse misure correttive. La più drastica è il rifiuto del contenuto o il blocco dell'utente che lo ha condiviso (se il controllo avvenisse a monte della pubblicazione) o la cancellazione o espulsione, qualora la moderazione sia successiva. Esistono misure più lievi, che sono generalmente le più praticate nel caso delle *fake news*. Esse includono: la modifica dei contenuti; la richiesta di chiarimenti; il commento o l'aggiunta di ulteriori informazioni, come il link alla fonte; la priorità, cioè la promozione o riduzione dell'accesso a un contenuto, ad esempio presentandolo in fondo alle ricerche, oppure escludendolo dai contenuti più rilevanti.

I rischi nella moderazione automatica delle fake news

Gli attuali sistemi di moderazione, specialmente quelli basati su tecniche di intelligenza artificiale, possono contribuire enormemente al contenimento delle *fake news*. Tuttavia, il loro utilizzo presenta rischi che possono portare all'esclusione indebita di contenuti di valore, e influire sulla libertà di espressione, l'accesso alle informazioni e il dialogo democratico.
Infatti, i sistemi di moderazione automatica si fondano generalmente su modelli probabilistici che prevedono sempre un margine di errore. Gli errori possono consistere in falsi positivi o falsi negativi: in caso di falso positivo, il sistema classifica come *fake news* una notizia veritiera, dunque precludendone la sua divulgazione o condivisione. Un falso negativo corrisponde a una *fake news* che viene classificata come veritiera quando in verità dovrebbe essere rimossa.
Un elemento di complicazione è poi determinato dall'ambiguità della stessa nozione di *fake news*. I sistemi di moderazione agiscono sulla base una rappresentazione computabile di una *fake news*. Ciò implica che lo sviluppatore del sistema abbia preventivamente definito

cosa sia una *fake news* e come sia possibile identificarla.
Esistono già molti studi che cercano di fornire una definizione del concetto di *fake news*, una formalizzazione o un *dataset* di riferimento, ma una definizione condivisa non è ancora disponibile. Per altro verso, non è sempre possibile avere accesso alla nozione di *fake news* implementata nel sistema.
Questo di solito accade con i sistemi di moderazione utilizzati da piattaforme che non diffondono la modalità operative dei sistemi (che altrimenti sarebbe facile da aggirare). Dunque, è bene tenere a mente che le decisioni automatiche riflettono sempre e necessariamente un giudizio arbitrario dello sviluppatore.
Un'ulteriore complessità è legata al fatto che l'identificazione di una *fake news* dipende spesso dall'intenzione del suo divulgatore e dal suo contesto sociale. Ad esempio, l'identità dell'autore (un comico o politico, un utente del servizio o concorrente), così come elementi contestuali (tempo e luogo) possono essere fattori dirimenti per determinare lo scopo e i possibili effetti di una certa modalità espressiva. Nonostante lo straordinario potere computazionale e la capacità di comprensione simbolico-sintattica, oggi i sistemi di intelligenza artificiale mostrano ancora forti limiti nella comprensione semantica del discorso e nel rilevare l'intenzione dietro le parole.
Il problema riguarda anche il trattamento giuridico di una *fake news.* Vi sono Paesi in cui la legislazione sulla libertà di espressione è più orientata verso garantire maggiore libertà di esprimersi, anche sostenendo tesi che potrebbero risultare come false (ad esempio negli Stati Uniti), e altri Paesi che invece ammettono una restrizione più incisiva laddove siano in gioco interessi di ordine pubblico (si pensi ad esempio alla divulgazione di *fake news* sul virus durante il periodo della pandemia Covid-19). Queste diversità nella cultura giuridica sono difficilmente riproducibili all'interno di sistemi che, eventualmente sviluppati sulla base di quadri normativi nazionali, si trovino poi a operare in contesti giuridici diversi.
Per via di queste complicazioni, è cruciale che i sistemi di moderazione delle *fake news* ammettano un certo grado di trasparenza e contestabilità.
Prima di tutto, il sistema dovrebbe consentire al soggetto che abbia condiviso la notizia di conoscere la misura adottata dal sistema di moderazione sulla presunta *fake news*.

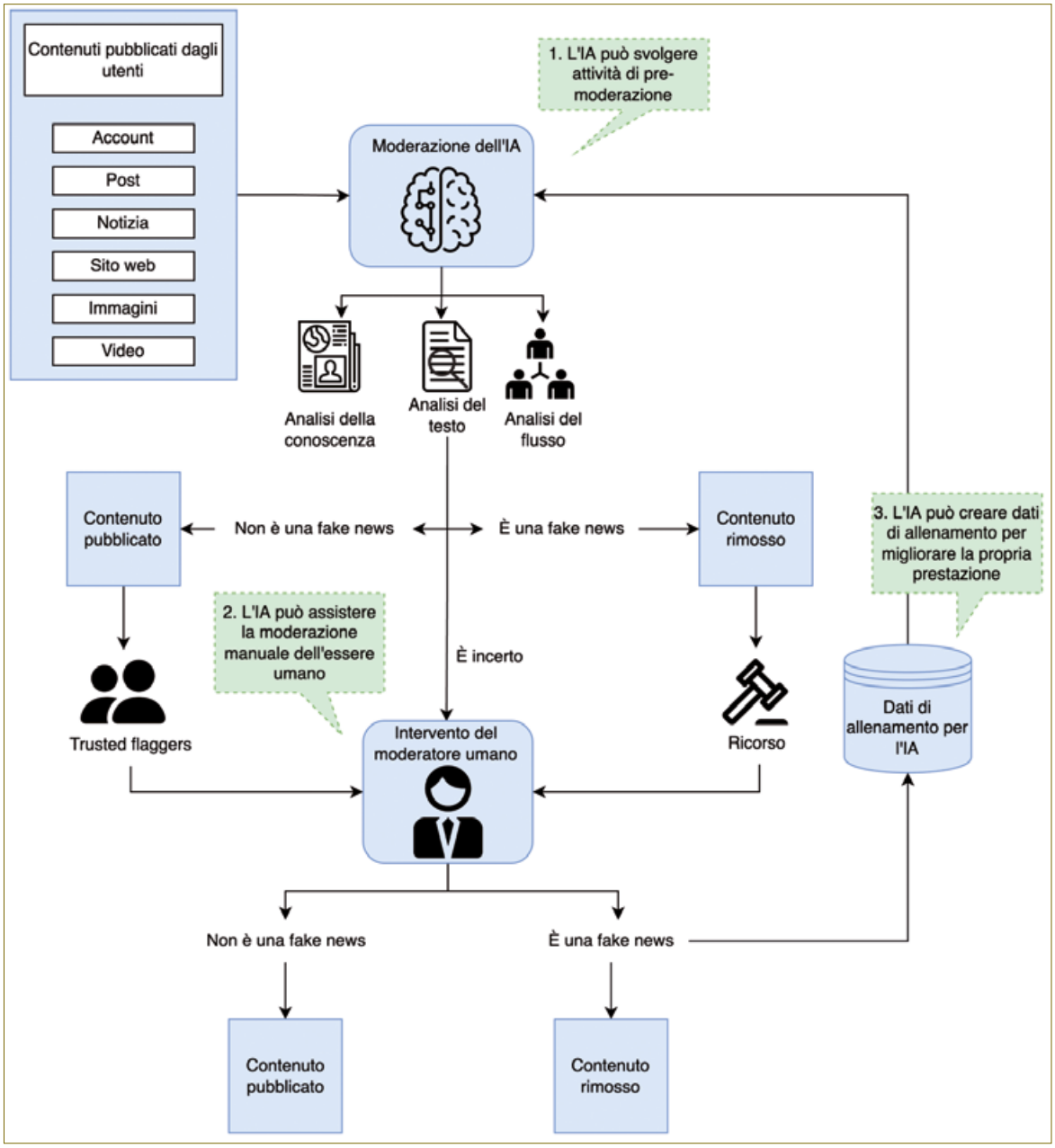

FIG. 1
FAKE NEWS

Funzionamento di un sistema complesso di moderazione.

Il sistema dovrebbe poi fornire una spiegazione del perché abbia classificato una notizia come *fake news*. Quando è possibile ricondurre la misura a una valutazione basata su fatti accertati, ciò dovrebbe includere il riferimento a tali fatti. Più difficile fornire una spiegazione, invece, quando l'individuazione delle *fake news* si basa su un'analisi linguistica e dello stile comunicativo del suo divulgatore. Ad esempio, una spiegazione che la notizia pubblicata ha utilizzato un certo numero di parole emotive o una certa sequenza di caratteri potrebbe non essere una spiegazione accettabile. Questo problema si ricollega all'incapacità degli attuali sistemi di Nlp di avere una reale comprensione dei significati delle parole e della connessione tra linguaggio e realtà sociale.
Vista le difficoltà di individuare con precisione una *fake news*, occorrerebbe poi dare sempre all'utente che abbia divulgato la notizia la possibilità di contestare la decisione del sistema e di dimostrarne la veridicità (ad esempio indicando fonti terze) oppure confutandone la falsità (adducendo argomenti).
Infine, quantomeno nei casi più dubbi, occorrerebbe sempre poter richiedere l'intervento di un essere umano esperto che possa giudicare.

Verso una governance della moderazione automatica

Da qualche anno ormai i problemi attinenti ai sistemi di moderazione sono entrati nel dibattito politico a livello Ue. Da una parte, si riconoscono le grandi potenzialità connesse all'utilizzo dei sistemi automatici per contribuire a governare il fenomeno delle *fake news*. Dall'altra si discutono misure adeguate a evitare effetti pregiudizievoli sulla libertà di espressione.
In particolare, sono state avviate molte iniziative di autoregolamentazione, non dotate di vincolatività giuridica, per garantire la trasparenza nella moderazione dei contenuti. Possiamo ricordare, ad esempio, i *Principi di Santa Clara*[3] approvati nel 2018, secondo i quali le imprese impegnate nella moderazione[4] automatica dei contenuti dovrebbero, tra le altre cose, informare gli utenti delle misure adottate sui contenuti e ammettere la possibilità di ricorso tempestivo.
In Ue, nel 2018, è stato istituito un codice di condotta per le piattaforme online e i principali operatori del settore pubblicitario per contrastare la disinformazione. Nel 2022, il codice è stato rinforzato inserendo importanti novità sull'utilizzo dei sistemi di

moderazione, tra cui assicurare che essi siano affidabili, rispettino i diritti degli utenti e non diano luogo a pratiche manipolatorie del dialogo e del comportamento degli utenti.
Oltre a queste iniziative, è in corso ormai da tempo una riflessione a livello europeo sull'opportunità di introdurre nuove regole per contrastare il fenomeno della condivisione di contenuti illeciti o antisociali online, tra cui anche le *fake news*. Il tema è legato alla necessità di ripensare in senso più rigoroso il regime di responsabilità dei prestatori intermediari per la diffusione di contenuti illeciti in rete. In questa direzione, la direttiva sul commercio elettronico del 2000, che prevedeva un ampio esonero della responsabilità degli intermediari, sarà sostituita dalla nuova legge sui servizi digitali, meglio nota come *Digital services act* (Dsa). Oltre a prevedere regole precise per le piattaforme online per tutelare i diritti fondamentali degli utenti in rete, il Dsa introduce regole dettagliate sull'utilizzo dei sistemi di moderazione. In particolare, vengono introdotte misure per garantire sia la trasparenza sia la contestabilità.
Riguardo la prima, gli intermediari devono includere nelle loro condizioni di servizio informazioni su eventuali restrizioni applicate nell'uso del servizio e sul funzionamento del sistema di moderazione. In secondo luogo, devono rendere disponibile, almeno una volta all'anno, un resoconto sulle attività di moderazione con informazioni specifiche come il numero dei contenuti rimossi classificati per tipo di violazione. Infine, gli intermediari che forniscono servizi di *hosting* devono sempre accompagnare la misura restrittiva imposta al contenuto con una motivazione chiara e specifica.
Riguardo la contestabilità, il Dsa prevede misure sia preventive sia successive per fornitori di servizi di *hosting* e per le piattaforme online. Da un lato, essi devono dotarsi di un meccanismo che consenta agli utenti di notificare la presenza di contenuti illegali. Dall'altro, debbono fornire accesso a un sistema interno di gestione dei reclami che consenta a ogni utente di presentare reclami elettronicamente e gratuitamente contro la misura di restrizione sul contenuto.
Quest'ultima appare necessaria per creare un ecosistema digitale che risponda alle esigenze degli utenti della piattaforma e ai valori sociali. È necessario però considerarne i limiti, incluso la fallibilità e parziarietà, e offrire tutele giuridiche adeguate. Il Dsa costituisce il primo passo nella direzione di una *governance* dei sistemi di moderazione automatica delle *fake news*.

Federico Galli[1], Andrea Loreggia[2], Giovanni Sartor[3]

1. Ricercatore, Università di Bologna
2. Ricercatore, Università di Brescia
3. Professore ordinario dell'Università di Bologna e dell'Istituto universitario europeo

NOTE

[1] https://transparency.fb.com/data/community-standards-enforcement/fake-accounts/facebook

[2] www.tripadvisor.com/TransparencyReport2021

[3] https://santaclaraprinciples.org/

[4] https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation

[5] https://eur-lex.europa.eu/legal-content/IT/LSU/?uri=CELEX:32000L0031

[6] www.agendadigitale.eu/mercati-digitali/digital-services-act-cose-e-cosa-prevede-la-legge-europea-sui-servizi-digitali/